1-30-2012

Bootstrapping for Significance of Compact Clusters in Multidimensional Datasets

Ranjan Maitra
Iowa State University, maitra@iastate.edu

Volodymyr Melnykov
North Dakota State University--Fargo, volodymyr.melnykov@ndsu.edu

Soumendra N. Lahiri
Texas A & M University - College Station

Bootstrapping for Significance of Compact Clusters in Multidimensional Datasets

Abstract
This article proposes a bootstrap approach for assessing significance in the clustering of multidimensional datasets. The procedure compares two models and declares the more complicated model a better candidate if there is significant evidence in its favor. The performance of the procedure is illustrated on two well-known classification datasets and comprehensively evaluated in terms of its ability to estimate the number of components via extensive simulation studies, with excellent results. The methodology is also applied to the problem of *k*-means color quantization of several standard images in the literature and is demonstrated to be a viable approach for determining the minimal and optimal numbers of colors needed to display an image without significant loss in resolution. Additional illustrations and performance evaluations are provided in the online supplementary material.

Keywords
Hierarchical clustering, k-means algorithm, Overlap, Prohorov metric, p-value quantitation map, q-value quantitation map

Disciplines
Statistics and Probability

Comments
This is an Accepted Manuscript of an article published by Taylor & Francis in *Journal of the American Statistical Association* on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

Bootstrapping for Significance of Compact Clusters in Multi-dimensional Datasets

Ranjan Maitra, Volodymyr Melnykov and Soumendra N. Lahiri *

A bootstrap approach is proposed for assessing significance in the clustering of multi-dimensional datasets. The developed procedure compares two models and declares the more complicated model a better candidate if there is significant evidence in its favor. The performance of the procedure is illustrated on two well-known classification datasets and comprehensively evaluated in terms of its ability to estimate the number of components via extensive simulation studies, with excellent results. The methodology is also applied to the problem of k-means color quantization of several standard images in the literature, and demonstrated to be a viable approach for determining the minimal and optimal numbers of colors needed to display an image without significant loss in resolution.

Keywords: bootstrap, overlap, hierarchical clustering, k-means algorithm, Prohorov metric, p-value quantitation map, q-value quantitation map

1. INTRODUCTION

Cluster analysis partitions datasets into groups such that observations within each class share common properties, but are also considerably different from those in other groups. This is a difficult problem with many diverse applications *eg*, taxonomical classification (Michener and Sokal, 1957), market segmentation (Hinneburg and Keim, 1999) or software management (Maitra, 2001) with consequently, a large body of literature in statistics and machine learning (see *eg.* Murtagh, 1985; Ramey, 1985; McLachlan and Basford, 1988; Kaufman and Rousseuw, 1990; Everitt et al., 2001; Fraley and Raftery, 2002; Tibshirani and Walther, 2005; Kettenring, 2006; Xu and Wunsch, 2009).

One popular approach to clustering is model-based: here, a probability model specifies the distribution of an observation in terms of a mixture of parametric distributions (Fraley and Raftery, 2002), with each mixture component describing properties of a particular group in the dataset. We refer to Titterington et al. (1985) and McLachlan and Peel (2000) for details, but note that clustering in such contexts requires a final assignment step that allocates each observation to the group for which its posterior probability of classification is highest.

The other main approach to clustering is largely heuristic and model-free. Prominent among these are the distance-based methods, with a distance measure between every pair of observations (or groups of observations). These algorithms are further sub-divided into the hierarchical (Johnson, 1967; Everitt et al., 2001; Jain and Dubes, 1988) and the partitional clustering algorithms which optimize some target function for a pre-specified number of clusters. Examples of the latter range from the classical k-means (Forgy, 1965; MacQueen, 1967) and k-medoids (Kaufman and Rousseuw, 1990) algorithms to the more modern approaches of competitive learning (Rumelhart and Zipser, 1985) or kernel-based clustering (Haykin, 1999; Xu and Wunsch, 2009).

An issue defying researchers in clustering is that of quantifying significance in the obtained groupings. Significance assessment in the clustering context itself needs to be defined: we address this in terms of whether a more complicated model (in terms of numbers of clusters, variables, estimable quantities, etc) is significantly better, in terms of its fit to a dataset, than a simpler model or whether any improvement can be explained purely as a matter of chance. We develop methodology in this paper for the case of the model-free clustering methods using the bootstrap.

Closely related to our look at significance assessment is that of estimating the number of clusters (in this paper, K) in a dataset: this long-standing issue (Everitt, 1979) has many suggested approaches (Milligan and Cooper, 1985; McLachlan and Peel, 2000). Suggestions in the distance-based case include Marriott (1971)'s criterion and its variants and the popular Gap statistic (Tibshirani et al., 2003). McLachlan (1987) suggests some approaches for model-based clustering that successively test for K against $K^*(> K)$ components, stopping when the null hypothesis can no

*Ranjan Maitra is Professor in the Department of Statistics and Statistical Laboratory, Iowa State University, Ames, IA 50011-1210, Volodymyr Melnykov is Assistant Professor in the Department of Statistics, North Dakota State University, Fargo, ND 50373, and Soumendra N. Lahiri is Professor in the Department of Statistics at Texas A& M University. This research was supported in part by the National Science Foundation CAREER Grant # DMS-0437555. The authors thank the Editor, an Associate Editor and three reviewers whose thoughtful comments greatly improved the content and presentation of this article.

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

longer be rejected, but development here has been muted by the need to account for possible regularity condition violations (Cramer, 1946). Challenges in implementation notwithstanding however, testing-based approaches provide at least one major advantage over the others, in that the results can be quantified in terms of the p-value, which is an universally understood measure between 0 and 1. Other approaches, while also quantitative, provide numerical values that are data- and context-dependent. For instance, Marriott's criterion which penalizes the logarithm of the determinant of the within-sums-of-squares-and-products matrix (W) by twice the log of the number of clusters provides a numerical value: however, the magnitude changes from one dataset to the other. There is thus no context-free and readily-interpretable quantification of the derived values, with no clear guidance on what differences in values are large and what are not. It is this more-interpretable quantification of the p-value that encourages us to give testing-based approaches another look at quantifying significance.

There has been some recent work in assessing significance in the derived clusterings. McShane et al. (2002) and Liu et al. (2008) use parametric mixture model assumptions and the bootstrap to individually test whether each of the derived groups can be further sub-divided into two, with allowances for high-dimensional low sample size (HDLSS) datasets, which forms the sole focus of their papers. (The use of the bootstrap in clustering is, however, of less recent vintage, having been employed (Kerr and Churchill, 2001; Dudoit and Fridlyand, 2003) to improve the reliability of clustering algorithms.) More recently, Maitra and Melnykov (2010a) derived an approximate approach for significance assessment in mixture models and model-based clustering. They developed a *quantitation map* which provides a researcher with a detailed quantitative measure summarizing evidence against simpler models and in favor of more complicated alternatives. Their derivations are however inapplicable in the context of more heuristic and parametric-model-free clustering methods.

In this paper, we provide a bootstrap approach for quantifying such clustering methods. We assume that we have compact groups which after some local transformation are spherical and similar to the other clusters. Under this framework, we develop in Section 2 a distribution-free bootstrap strategy for testing a particular clustering setup vis-a-vis a more complicated one, *i.e.*, we test between a null K- and an alternative K^*-clusters model. To fix ideas for the development of our methodology here, we assume that $K^* > K$ and that data partitions with more clusters are somehow always more complicated than those with fewer groups. We emphasize that this assumption is solely for expediency and ease of presentation in this paper: our methodology also applies to other scenarios. The performance of our approach is illustrated on two well-known classification datasets in Section 3 and evaluated in terms of its ability to estimate the number of significant groups through a series of simulation experiments in Section 4. The consistency of the bootstrap methodology for our problem is theoretically explored in Section 5. We next illustrate utility of our methodology in Section 6 to the novel application of determining the number of colors required to adequately represent a digital image. This paper concludes with some discussion in Section 7 along with an outline of some possible directions for future work. We also have a supplement providing some additional illustrations and performance evaluations. Sections, figures and tables in the supplement referred to in this paper are labeled with the prefix "S-".

2. METHODOLOGY

2.1 Background and Preliminaries

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample of n p-dimensional observations. We assume that each $\boldsymbol{X}_i \sim \sum_{k=1}^{K} \zeta_{ik} f_k(\boldsymbol{x})$, where K is the number of groups, $\zeta_{ik} = \mathcal{I}_{(\boldsymbol{X}_i \in \mathcal{G}_k)}$, and f_k is the density of an observation in the kth cluster. Here $\mathcal{I}_{(\cdot)}$ is the indicator function and $\mathcal{G}_k$ is the set of observations in the sample belonging to the k-th cluster. The primary objective is to estimate K and ζ_{ik}s for $i = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, K$ in the presence of the nuisance parameters $f_k(\cdot)$. We assume that for each $k = 1, 2, \ldots, K$, there exists a function $\Psi_k : I\!R^p \rightarrow I\!R^p$ such that $g(\boldsymbol{x}) = f_k(\Psi_k(\boldsymbol{x}))|J_{\Psi_k}|$ where $|J_{\Psi_k}|$ is the Jacobian of Ψ_k and $g(\boldsymbol{x})$ is a density centered at the origin with spherical level hyper-surfaces, *i.e.* for any orthogonal $p \times p$-matrix $\boldsymbol{\Gamma}$, $g(\boldsymbol{x}) = g(\boldsymbol{\Gamma x})$. Further $g(\boldsymbol{y}) = \prod_{j=1}^{p} h(y_j)$ for $\boldsymbol{y} = (y_1, y_2, \ldots, y_p)$ where $h(\cdot)$ is an univariate density symmetric about zero. This is the most general formulation, but in this paper, we consider only those cases which result in center-based clusterings. For instance, when $\Psi_k(\boldsymbol{x}) \equiv \boldsymbol{x} - \boldsymbol{\mu}_k$, then each group has the same density but for location given by $\boldsymbol{\mu}_k$ and we essentially assume that we have homogeneous spherical clusters. This is the putative framework underlying the k-means and other Euclidean distance-based clustering algorithms. When $\Psi_k(\boldsymbol{x}) \equiv \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{x} - \boldsymbol{\mu}_k)$, we assume having ellipsoidal clusters of different shapes and orientations centered at the $\boldsymbol{\mu}_k$s and Mahalanobis' distance-based clustering methods would be most appropriate.

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

2.2 A Hypothesis-testing Framework

Suppose we want to investigate if there is significant evidence that the given dataset is better described by K^* compact groups than by K clusters, where $K^* > K$. The null and alternative hypotheses can then be prescribed in the form $H_0 : F \in \mathcal{F}_K$ vs. $H_a : F \in \mathcal{F}_{K^*}$,where F represents the true distribution of the observations in the dataset, and $\mathcal{F}_K$ and $\mathcal{F}_{K^*}$ represents the family of distributions under the null and alternative hypothesis respectively. The ith observation has density of the form $\sum_{k=1}^{K} \zeta_{ik}^{(K)} f_k^{(K)}(\boldsymbol{x})$ under H_0 and $\sum_{k=1}^{K^*} \zeta_{ik}^{(K^*)} f_k^{(K^*)}(\boldsymbol{x})$ under H_a.

2.2.1 Test statistic

This paper uses as the test statistic the improvement in the within-cluster-sum-of-squares $s_{K;K^*} = W_K - W_{K^*}$, where $W_K = \sum_{i=1}^{n} \sum_{k=1}^{K} \zeta_{ik}^{(K)} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)'(\boldsymbol{x}_i - \boldsymbol{\mu}_k)$, the optimized objective function for terminated k-means algorithms that is obtained as its by-product when used to group a dataset into K clusters. Note, however, that while we develop methodology here using $s_{K;K^*}$ as our test statistic, our development is general enough to also extend to other reasonable test statistics. Note also that $s_{K;K^*} \geq 0$ always, since W_k decreases as K increases: our objective is to assess the significance in the improvement of W_K upon fitting the dataset with a K^*-cluster model over that with only K components. To do so, we have to calculate the p-value of the test statistic in terms of the probability that a $s_{K;K^*}$ from a true K-cluster model is greater than the $s_{K;K^*}$ calculated from the dataset. We develop approaches to estimating this p-value next.

2.3 Obtaining a reference distribution

2.3.1 Homogeneous spherical clusters

We first develop methodology for homogeneous clusters. Under the null hypothesis of K groups, we have a sample $\boldsymbol{\Xi} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ from the joint distribution

$$\prod_{i=1}^{n} \sum_{k=1}^{K} \zeta_{ik}^{(K)} \frac{1}{\sigma} g\left(\frac{x_i - \mu_k}{\sigma}\right), \tag{1}$$

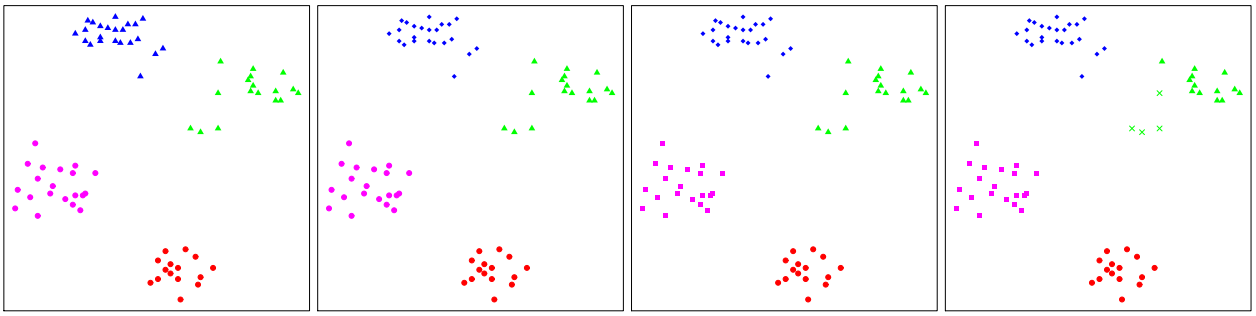
where $g : I\!R^p \to I\!R^p$ is invariant under orthogonal transformations. Our objective is to obtain the null distribution of $s_{K;K^*}$ under the assumption that the data are realizations from (1).

In obtaining the null distribution, we note that μ_ks, $\zeta_{ik}^{(K)}$s, σ and $g(\cdot)$ are all parameters of our null distribution. We use the K-clusters solution to the dataset to obtain estimates $\hat{\zeta}_{ik}^{(K)}$s, $\hat{\mu}_k = \sum_{i=1}^{n} \hat{\zeta}_{ik}^{(K)} X_i / \sum_{i=1}^{n} \hat{\zeta}_{ik}^{(K)}$ and $\hat{\sigma} = \frac{1}{np} \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{\zeta}_{ik}^{(K)} (X_i - \hat{\mu}_k)'(X_i - \hat{\mu}_k)$. Note that once the effects of the assigned centers $\hat{\boldsymbol{\mu}}_k$ and common scale $\hat{\sigma}$ have been removed from each observation (using the K-clusters solution), then the residuals $\hat{\boldsymbol{\epsilon}}_1, \hat{\boldsymbol{\epsilon}}_2, \ldots, \hat{\boldsymbol{\epsilon}}_n$, where $\hat{\boldsymbol{\epsilon}}_i = (\boldsymbol{X}_i - \sum_{k=1}^{K} \hat{\zeta}_{ik}^{(K)} \hat{\boldsymbol{\mu}}_k)/\hat{\sigma}$ for $i = 1, 2, \ldots, n$, form a sample from the common density $g(\cdot)$. A naïve approach to estimating $g(\cdot)$ and using the bootstrap would resample from these $\hat{\boldsymbol{\epsilon}}_i$s. This naïve approach would, specifically, involve obtaining $\hat{\boldsymbol{\epsilon}}_1^\star, \hat{\boldsymbol{\epsilon}}_2^\star, \ldots, \hat{\boldsymbol{\epsilon}}_n^\star$ by sampling with replacement from $\hat{\boldsymbol{\epsilon}}_1, \hat{\boldsymbol{\epsilon}}_2, \ldots, \hat{\boldsymbol{\epsilon}}_n$, and then constructing the resampled realizations $\boldsymbol{X}_i^\star = \sum_{k=1}^{K} \hat{\zeta}_{ik}^{(K)} \hat{\boldsymbol{\mu}}_k + \boldsymbol{\epsilon}_i^\star, i = 1, 2, \ldots, n$. Theorem 5.1 shows that such a resampling strategy has low power, especially in the case when H_0 specifies far fewer groups than the true. Additional illustration of these claims is provided in the supplemental file: in particular, see Figure S-1. Thus, a carefully-designed resampling strategy is needed.

Note that $g(\cdot)$ is assumed to have spherical level hyper-surfaces and is defined through the univariate density $h(\cdot)$. Given this special structure for $g(\cdot)$, we note that the ordered set of normed residuals $\|\hat{\boldsymbol{\epsilon}}_{(1)}\| \leq \|\hat{\boldsymbol{\epsilon}}_{(2)}\| \leq \ldots \leq \|\hat{\boldsymbol{\epsilon}}_{(n)}\|$ is sufficient for $g(\cdot)$. Thus, the conditional distribution of $\hat{\boldsymbol{\epsilon}}_1, \hat{\boldsymbol{\epsilon}}_2, \ldots, \hat{\boldsymbol{\epsilon}}_n$ given the ordered set of normed residuals $\|\hat{\boldsymbol{\epsilon}}_{(1)}\|, \|\hat{\boldsymbol{\epsilon}}_{(2)}\|, \ldots, \|\hat{\boldsymbol{\epsilon}}_{(n)}\|$ is free of $g(\cdot)$. We use this fact to obtain our resampled realizations. Note also that transforming $\hat{\boldsymbol{\epsilon}}_i$ to $\boldsymbol{\Gamma}_i \hat{\boldsymbol{\epsilon}}_i, i = 1, 2, \ldots, n$, where $\boldsymbol{\Gamma}_i$ is any $p \times p$ orthogonal matrix, maintains the ordered set of normed residuals $\|\hat{\boldsymbol{\epsilon}}_{(1)}\|, \|\hat{\boldsymbol{\epsilon}}_{(2)}\|, \ldots, \|\hat{\boldsymbol{\epsilon}}_{(n)}\|$. Thus, a resampling strategy for the residuals would involve randomly sampling p-dimensional unit direction vectors, scaling them first with the magnitude of the residuals $\|\hat{\boldsymbol{\epsilon}}_{(i)}\|$ and then with $\hat{\sigma}$ and shifting by the appropriate $\hat{\boldsymbol{\mu}}_k$ corresponding to $\hat{\zeta}_{ik}^{(K)}$.

Formally therefore, we propose the following: for each $i = 1, 2, \ldots, n$, first generate a random unit direction in p-dimensional space. We do so by simulating a p-variate standard normal random vector $\boldsymbol{Z}_i$, and let $\boldsymbol{W}_i = \boldsymbol{Z}_i / \|\boldsymbol{Z}_i\|$. Note that we use a standard p-variate Gaussian distribution random to obtain our normed realization, but

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.



(a) $K = 2, \mathcal{R} = 0.5$ (b) $K = 3, \mathcal{R} = 0.746$ (c) $K = 4, \mathcal{R} = 1.0$ (d) $K = 5, \mathcal{R} = 0.958$

Figure 1. Ruspini dataset: k-means clustering solutions for different K. In each figure, colors represent true clusterings while characters represent the predicted groupings.

any distribution spherically symmetric around zero would also be appropriate, following Theorem 1 of Cambanis et al. (1981). We next obtain a random permutation $(\ell_1, \ell_2, \ldots, \ell_n)$ of $\{1, 2, \ldots, n\}$. Then, the ith resampled residual is given by $\boldsymbol{\epsilon}_i^* = \|\hat{\boldsymbol{\epsilon}}_{\ell_i}\| \boldsymbol{W}_i$. Note that $\|\boldsymbol{\epsilon}_i^*\| = \|\hat{\boldsymbol{\epsilon}}_{\ell_i}\| \|\boldsymbol{W}_i\| = \|\hat{\boldsymbol{\epsilon}}_{\ell_i}\| \|\boldsymbol{Z}_i\| / \|\boldsymbol{Z}_i\| = \|\hat{\boldsymbol{\epsilon}}_{\ell_i}\|$, so that the resampled residuals in the set $\{\boldsymbol{\epsilon}_i^*; i = 1, 2, \ldots, n\}$ maintain the norms in the set $\{\boldsymbol{\epsilon}_i; i = 1, 2, \ldots, n\}$. (The directions of $\boldsymbol{\epsilon}_i^*$s are however different from those in $\hat{\boldsymbol{\epsilon}}_i$s, following the effect of $\boldsymbol{W}_i$.) Adding these $\hat{\boldsymbol{\epsilon}}_i^*$s, after scaling with $\hat{\sigma}$, to the means of the corresponding cluster centers yields $\boldsymbol{X}_i^* = \sum_{k=1}^K \hat{\zeta}_{ik}^{(K)} \hat{\boldsymbol{\mu}}_k + \hat{\sigma} \boldsymbol{\epsilon}_i^*$ for each $i = 1, 2, \ldots, n$. Thus, we get a resampled realization of the dataset under H_0: $\boldsymbol{\Xi}^* = \{\boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \ldots, \boldsymbol{X}_n^*\}$. We replicate the procedure M times to obtain resampled realizations of the dataset $\boldsymbol{\Xi}_1^*, \boldsymbol{\Xi}_2^*, \ldots, \boldsymbol{\Xi}_M^*$. From each $\boldsymbol{\Xi}_j^*$, we obtain the test statistic given by $s_{j,(K;K^*)}^*$, for $j = 1, 2, \ldots, M$. The p-value of $s_{K;K^*}$ is then estimated by the proportion of cases in which it is exceeded by the resampled $s_{j,(K;K^*)}^*$. Formally therefore, the p-value of the test statistic is calculated from $\frac{1}{M} \sum_{j=1}^M \mathcal{I}\left(s_{j,(K;K^*)}^* > s_{K;K^*}\right)$, where $\mathcal{I}(\cdot)$ is again the indicator function.

An Illustration We illustrate performance of our resampling mechanism on the synthetic Ruspini (1970) dataset which contains 75 bivariate observations from four well-separated groups that are each fairly spherical in their spread. This dataset, also available in the CLUSTER package in R, is popular for investigating clustering algorithms (Struyf et al., 1997). Figure 1 shows the optimal 2-, 3-, 4- and 5-cluster solutions obtained using the k-means algorithm, initialized here – as in all experiments reported in this paper – using the best (in terms of the smallest W_K) of the deterministic approach of Maitra (2009) and the partitioning obtained by using the hierarchical clustering algorithm with Ward's linkage. For each clustering, we also report the *Adjusted Rand* index ($\mathcal{R}$) (Hubert and Arabie, 1985) which measures similarity between two partitionings – in this case, the derived grouping and the true. (Note that $\mathcal{R}$ takes its maximum value of 1 when the two partitionings match perfectly. In general, values of $\mathcal{R}$ close to unity indicate good clustering performance while those far below 1 are indicative of poorer performance.) In testing for the adequacy of a 1-, 2-, 3-cluster solution vis-a-vis a solution with (say) 4 clusters, we need resampled datasets from the corresponding null distribution. Figures S-2 (in the supplemental file) provide four resampled datasets each under null hypothesis assumptions of 1, 2, 3 and 4 clusters. The realizations for the first three sets appear quite different from the Ruspini data so that any reasonable test statistic should have low probability of accepting H_0 when challenged by a 4-group model. On the other hand, when testing under a H_0 of four true groups in the dataset, resampled datasets (refer again to Figures S-2) look quite similar to that of the observed data so that the test statistic will have a lower chance of rejecting H_0. We return to this dataset a little later in Section 3, moving instead to generalizing our methodology for cases beyond homogeneous spherical clusters.

2.3.2 Extension to the case of general ellipsoidal clusters

The entire process can be easily adapted for the case of general ellipsoidal clusters. Under H_0, our sample $\boldsymbol{\Xi} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ is from the joint distribution with K groups given by

$$\prod_{i=1}^n \sum_{k=1}^K \zeta_{ik}^{(K)} \frac{1}{\det(\boldsymbol{\Sigma}_k)} g\left(\boldsymbol{\Sigma}_k^{-\frac{1}{2}} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right), \tag{2}$$

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

where $g : I\!R^p \rightarrow I\!R^p$ is as in (1). Once again, noting that $\boldsymbol{\mu}_k$s, $\zeta_{ik}^{(K)}$s, $\boldsymbol{\Sigma}_k$s and g are all parameters under H_0, we use the K-clusters solution to the dataset and obtain estimates $\hat{\zeta}_{ik}^{(K)}$s, $\hat{\boldsymbol{\mu}}_k = \sum_{i=1}^n \hat{\zeta}_{ik}^{(K)} \boldsymbol{X}_i / \sum_{i=1}^n \hat{\zeta}_{ik}^{(K)}$ and $\hat{\boldsymbol{\Sigma}}_k = \sum_{i=1}^n \hat{\zeta}_{ik}^{(K)} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}_k)' / \sum_{i=1}^n \hat{\zeta}_{ik}^{(K)}$. After the effects of the assigned center and individual scale have been removed from each observation, and writing each residual $\hat{\boldsymbol{\epsilon}}_i = \sum_{k=1}^K \hat{\zeta}_{ik}^{(K)} \hat{\boldsymbol{\Sigma}}_k^{-\frac{1}{2}} (\boldsymbol{X}_i - \sum_{k'=1}^K \hat{\zeta}_{ik'}^{(K)} \hat{\boldsymbol{\mu}}_{k'})$ for $i = 1, 2, \ldots, n$, we are left with the same scenario as in Section 2.3.1. That is, we have $\hat{\boldsymbol{\epsilon}}_1, \hat{\boldsymbol{\epsilon}}_2, \ldots, \hat{\boldsymbol{\epsilon}}_n$ from the common density g with spherical level hyper-surfaces and defined through the univariate density h. We obtain resampled residuals $\boldsymbol{\epsilon}_1^*, \boldsymbol{\epsilon}_2^*, \ldots, \boldsymbol{\epsilon}_n^*$ in the same manner as before. Combining, our resampled realization from the null distribution is given by $\boldsymbol{\Xi}^* = \{\boldsymbol{X}_1^*, \boldsymbol{X}_2^*, \ldots, \boldsymbol{X}_n*\}$, where $\boldsymbol{X}_i^* = \sum_{k=1}^K \hat{\zeta}_{ik}^{(K)} (\hat{\boldsymbol{\mu}}_k + \hat{\boldsymbol{\Sigma}}_k^{\frac{1}{2}} \boldsymbol{\epsilon}_i^*)$ for each $i = 1, 2, \ldots, n$. As before, the procedure is replicated M times to obtain resampled realizations $\boldsymbol{\Xi}_1^*, \boldsymbol{\Xi}_2^*, \ldots, \boldsymbol{\Xi}_M^*$ under H_0, from each of which we get $s_{j,(K;K^*)}^*$, for $j = 1, 2, \ldots, M$. The p-value of the observed test statistic is again estimated by $\frac{1}{M} \sum_{j=1}^M \mathcal{I}\left(s_{j,(K;K^*)}^* > s_{K;K^*}\right)$.

We refer to the supplemental file for an illustrative example of the methodology developed here (see Section S-2 and, in particular, Figures S-3 and S-4). Note that while the development here has been for the case of general ellipsoidal clusters, it is also potentially valid for clusters that are more general-shaped, as long as an appropriate transformation $\boldsymbol{\Psi}$, *e.g.* the multivariate Box-Cox transform, can be found (see Section S-5 for an illustrative two-dimensional example). We also note that our methodology is applicable only to the case of multi-dimensional datasets (*i.e.*, $p \geq 2$). It is, by construction, unlikely to be able to generate a rich ensemble of realizations from the null distribution for one-dimensional observations, without additional assumptions.

2.3.3 Null distribution with no clusters

Sections 2.3.1 and 2.3.2 developed methodology for simulating realizations from the null distribution with K clusters. This methodology is however inapplicable for the case where the null distribution specifies no clustering in the data (a case which we refer to here as $K = 0$). This is a scenario where the null distribution is uniform over the support of the data, so we propose adopting Tibshirani et al. (2003)'s proposal of sampling from the uniform distribution on the p-dimensional hyper-rectangle (used by the authors to decide on the number of clusters for the Gap statistic).

2.4 Summarizing Significance via Quantitation Maps

Maitra and Melnykov (2010a) also developed the *p-value quantitation map* to provide comprehensive visualization of the p-values for different tests in the context of mixture models. The rows of these two-dimensional upper-triangular maps index the model under H_0 while the columns denote the model under H_a. The intersection of a particular row-column pair yields the p-value of the test statistic for testing the corresponding H_0 against the corresponding H_a. To address the issue of multiple significance, the authors proposed controlling for the expected *false discovery rates* (FDR) using Benjamini and Hochberg (1995), giving rise to the *q-value quantitation map*. We note that, as cautioned by a reviewer, these derived q-values are somewhat *ad hoc* because the simultaneous hypothesis tests being tested have to be approximately independent for the methodology of Benjamini and Hochberg (1995) to apply. Our detailed experiments in Sections 3 and 4, however, show that these quantitation maps work quite well in practice.

2.4.1 Application to choosing K

The p- and q-value quantitation maps allow us to assess the significance of a host of clustering solutions. One important application of these maps is to obtain an optimal estimate of K, from among a pre-specified range $[K_{\min}, K_{\max}]$. The quantitation map drawn represents p- and q-values with K corresponding to the simpler model in H_0 and the one (K^*) corresponding to the more complicated model as H_a. In this application, we assume that $K^* > K$, though this is not a constraining requirement – see Section 2.4.2. Maitra and Melnykov (2010a) suggest choosing the optimal K from within the range $[K_{\min}, K_{\max}]$ by sequential testing using the following algorithm:

1. Let $K = K_{\min}$ and $K^* = K_{\min} + 1$.
2. If the q-value for testing $H_0 : K$ versus $H_a : K^*$ is less than the desired FDR (q_0, say), reassign $K \leftarrow K + 1$ and $K^* \leftarrow K^* + 1$. Otherwise, reassign $K^* \leftarrow K^* + 1$.

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

3. Reiterate Step 2 until $K^* > K_{\max}$. Report the current (null) K as the number of clusters.

This scenario is less conservative than sequentially testing $H_0 : K$ versus $H_a : K+1$ clusters for $K = K_{\min}, K_{\min}+2, \ldots, K_{\max}-1$ until the first instance for which $q \geq q_0$. We prefer the above approach because the failure to detect significance for some $H_0 : K$ versus $H_a : K+1$ for some K does not necessarily rule out the possibility that another K^*-clustering solution for some $K^* > K+1$ would be significantly better than the K-solution.

2.4.2 Choosing between different clustering solutions

We conclude our discussion in this section by mentioning that the methodology in Section 2.4.1 can be readily generalized to make statements on the significance of many aspects of models. As a specific example, note that we can use the above development to identify whether a more general K^*-clusters model is significantly better than a less complicated model with K-groups, where complexity is determined based on a number of factors, *eg.* the number of parameters being estimated under each hypothesis. We illustrate this generalization in Section 3.2.1 where we investigate groupings of the Iris dataset assuming spherical and general-shaped clusters.

3. ILLUSTRATIVE EXAMPLES

3.1 The *Ruspini* dataset

Our first illustration is on the dataset of Section 2.3.1, clustered via k-means for different K. (For all cases in this paper, $M = 1000$.) The q-value quantitation map (Figure 2) indicates that any clustering solution ($K^* > 0$) is preferable over one with no clustering ($K = 0$), and that any of the K^*-group ($K^* > 1$) partitions is significantly better ($q < 0.05$) than assuming a homogeneous structure in the data ($K = 1$). Indeed, it appears that any of the partitions obtained using $K^* > K$ groups for $K^* \leq 7$ is also significantly better than the K-groups partition, for $K = 2, 3$, but the same can not be said for when $K = 4$. Thus, we are led to prefer the model with $K = 4$ as it is the first instance over which more complicated models are not significantly preferred. Indeed, for this solution, we get a perfect match ($\mathcal{R} = 1.0$), while $\mathcal{R} = 0.5$ and 0.746 for the 2- and 3-clusters solutions, respectively.

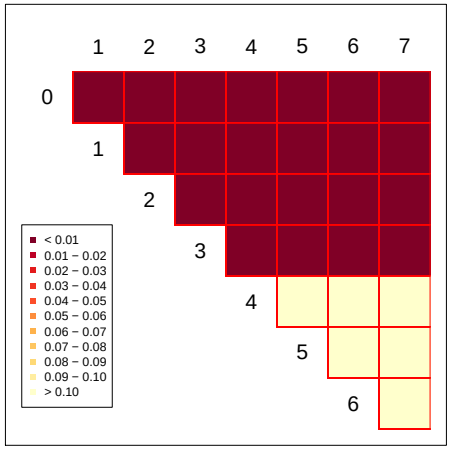


Figure 2. The q-value quantitation map for the Ruspini dataset.

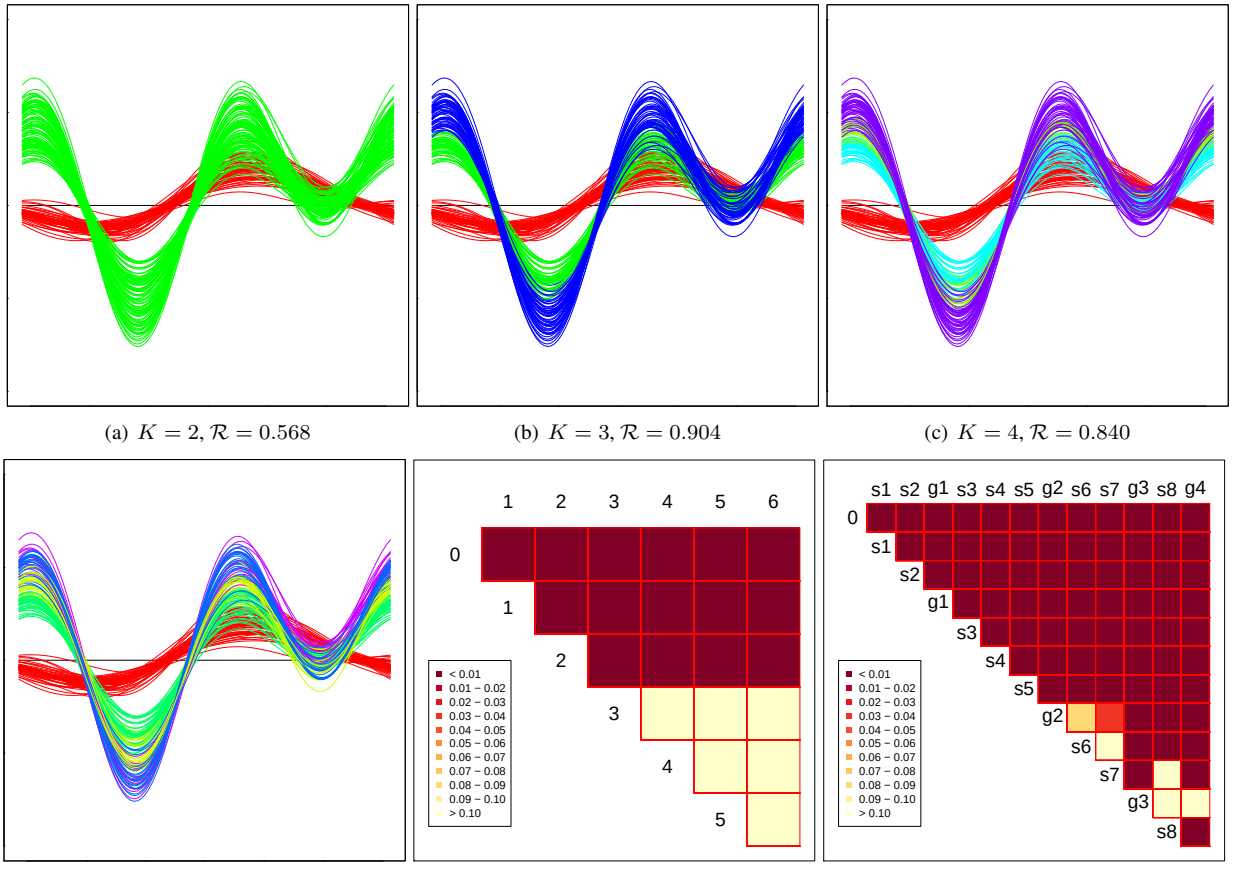
3.2 *Iris* dataset

Our next illustration is on the celebrated *Iris* dataset (Anderson, 1935; Fisher, 1936) having measurements on each of petal length and width and sepal length and width on 50 observations each drawn from three *Iris* species, namely *I. setosa*, *I. versicolor* and *I. virginica*. It is known that *I. setosa* is very clearly distinguished from the other two species while *I. versicolor* and *I. virginica* are more closely related and not as easily separated. We partition the dataset for different K using Gaussian model-based clustering with general dispersions – initialized using the *emEM* approach of Biernacki et al. (2003) – and then bootstrap for significance as per Section 2.3.2. Figures 3a–d present modified Andrews (1972) curves (formal description also provided in the supplemental file, vide Section S-3.1) for the 2-, 3-, 4- and 5-groups partitionings, along with their $\mathcal{R}$-values relative to the true group identifications. The distinctiveness of *I. setosa* from the other two species is very clear (supported by Figure S-3a of the supplement, which displays the true classification). The 3-group solution also shows some overlap between *I. versicolor* and *I. virginica*, while the 4- and 5- group solutions show these two species as being sub-divided further. Figure 3e displays the q-value quantitation map: the no clustering assumption is clearly not tenable. We also notice significant improvements upon fitting larger K^*-groups-models ($K^* \leq 6$) over models with one or two groups. (In particular, $q = 0.005$ for testing $H_0 : K = 2$ versus $H_a : K^* = 3$. The quantitation map also informs us that more complicated clustering solutions are not significantly preferred over the 3-cluster model ($q > 0.1$ in all cases), thus three groups are adequate to describe the heterogeneity in the data. This solution also most closely matches the known classification ($\mathcal{R} = 0.904$) for the dataset.

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

3.2.1 Choice of clustering algorithm

A reviewer wondered about the choice of mixtures-of-Gaussians-model-based clustering with general group-specific dispersions instead of, for instance, k-means clustering which inherently assumes a common spherical dispersion structure for all groups. This interesting question can be answered by the development of Section 2.4.2. Figure 3f provides the q-value quantitation map for comparing models fit using k-means (K ranging from 1 through 8) and model-based clustering (K varying from 1 through 4): here the models are ordered according to the number of parameters reflecting model complexity. Thus, the model with fewer parameters (Kp for k-means, $\binom{K+1}{2}p + Kp + K - 1$ for model-based clustering) is in H_0 (row) while that with the larger number of parameters is in H_a (column). Further supporting evidence in the form of Andrew's curves and $\mathcal{R}$-values of the k-means clustering solutions is in Figure S-5 of the supplement. Figure 3f provides us with an understanding of several aspects: for instance, the solution with seven homogeneous spherical groups is significantly better ($0.02 < q < 0.03$) than the solution that places the *I. setosa* observations in one cluster and the others in one other group. Also, if we restrict to only homogeneous spherical groups, no larger model fits the dataset significantly better ($q > 0.10$) than the 7-groups solution. (For clarity of presentation, Figure 3f does not display models with more than 8 homogeneous spherical groups.) But the 3-cluster model with general group-specific dispersions is our optimal choice, and most appropriately so, since its $\mathcal{R}$-value again indicates that it most closely fits the data.



(a) $K = 2, \mathcal{R} = 0.568$ (b) $K = 3, \mathcal{R} = 0.904$ (c) $K = 4, \mathcal{R} = 0.840$

(d) $K = 5, \mathcal{R} = 0.732$ (e) q-value quantitation map for models with non-spherical covariances (f) q-value quantitation map for various models

Figure 3. (a-d) Andrews' curves of Iris data colored according to the groupings obtained using Gaussian model-based clustering with general group-specific dispersions for 2, 3, 4 and 5 groups, respectively. (e) The corresponding q-value quantitation map. (f) The q-value quantitation map for models fit using k-means and model-based clustering with different K. The letters "s" and "g" in the row and column labels are for k-means-obtained and model-based-clustering-obtained groupings respectively, while the numerals indicate the number of groups.

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

4. EXPERIMENTAL EVALUATIONS

We performed extensive simulation experiments to evaluate performance of our methodology. Although not the only application of our algorithm, in order to facilitate comparisons with other methods, we evaluated performance by using the methodology of Section 2.4.1 to estimate K in datasets of many combinations of p, K and n, with both homogeneous spherical and nonhomogeneous ellipsoidal clusters. Datasets were obtained using the `R` package MIXSIM which provides a convenient way to simulate clustered data from Gaussian mixture models with pre-specified overlap characteristics as a surrogate for clustering complexity (Maitra and Melnykov, 2010b). These overlap measures are summarized in the form of the average ($\bar{\omega}$) and/or maximum ($\breve{\omega}$) of all pairwise overlaps, with larger values corresponding to greater clustering difficulty. In the general case (with nonhomogeneous ellipsoidal clusters) MIXSIM can simulate clustered data after controlling for both $\breve{\omega}$ and $\bar{\omega}$ yielding a wide range of cluster geometries (Maitra and Melnykov, 2010b). We therefore controlled both $\breve{\omega}$ and $\bar{\omega}$ for these cases. However, computational limitations in MIXSIM make it infeasible to generate datasets with homogeneous spherical clusters while controlling both $\breve{\omega}$ and $\bar{\omega}$, so then we only set $\breve{\omega}$. We obtained 100 simulated datasets at each combination of $(p, K, n, (\bar{\omega}, \breve{\omega}))$ for the general nonhomogeneous ellipsoidal clusters case, and at each combination of $(p, K, n, \breve{\omega})$ for the case with homogeneous spherical clusters. For all experiments, performance was compared with results obtained using the *gap statistic* proposed by Tibshirani et al. (2003) which estimates the number of clusters in datasets by comparing the change in observed within-cluster variation with the expected under a null (no-groups) model. Our implementation of the gap statistic used both the untransformed data (*Gap*) and the more commonly-used variant that applies it on the dataset transformed using its singular value decomposition (*GapSVD*). For each simulated dataset and method, we calculated $\mathcal{R}$ for the derived grouping (obtained at the estimated $\hat{K}$) relative to the true. We compared these values with the best possible $\mathcal{R}$ that could be obtained using the clustering algorithm: this was obtained by applying the corresponding clustering algorithm (k-means for the homogeneous spherical case, hierarchical clustering with Ward's criterion for the more general case) for each K, calculating $\mathcal{R}$ of the derived grouping relative to the true, and taking the maximum of these $\mathcal{R}$s. We define the difference between this maximum $\mathcal{R}$ and the $\mathcal{R}$ obtained from the grouping at the estimated $\hat{K}$ to be the adjusted Rand Index deficiency and denote this deficiency by $\mathcal{D}$. We evaluated performance of all methods in terms of these $\mathcal{D}$-values.

4.1 The Case with Homogeneous Spherical Clusters

In all experiments reported in this section, we partitioned each dataset for different K using k-means. In addition to Section 2.3.1, *Gap* and *GapSVD*, we also compared with Krzanowski and Lai (1985)'s approach (*KL*) which chooses $\hat{K} = \text{argmax}_k\{KL(k) : 2 \leq k \leq K_{\max}\}$, where

$$KL(k) = \left| \frac{(k-1)^{\frac{2}{p}} W_{k-1} - k^{\frac{2}{p}} W_k}{k^{\frac{2}{p}} W_k - (k+1)^{\frac{2}{p}} W_{k+1}} \right|.$$

Note that *KL* assumes that there is more than one group in the data: thus, it is constrained to only choose among models that have $\hat{K} > 1$ components.

4.1.1 Datasets with Compact Clusters

Our combinations of $(K, p, n, \breve{\omega})$ set $K = 6$ or 11, $p = 5$ or 10, $n = 500$ or 1000, and $\breve{\omega} = 0.01$, 0.05 or 0.01. Table 1 displays the mean and standard deviation of $\hat{K}$ (left block) and $\mathcal{D}$ (right block) summarized over the 100 simulated datasets at each setting for each methodology. Additional supporting evidence in the form of a graphical display of the distributions of $\hat{K}$'s and the $\mathcal{D}$s for each methodology at each setting is provided in the supplement – see Figure S-6. The table and the figure both indicate that performance of our proposed bootstrap-based procedure was slightly better than *Gap*, but both methods performed much better than either *GapSVD* and *KL*. Interestingly, *GapSVD* sometimes detected only one cluster, even in cases with low clustering difficulty but *Gap* and our bootstrap-based approach did not share this shortcoming.

4.1.2 Datasets with Heavier-tailed Clusters

We also investigated performance with heavier-tailed clusters than the Gaussian. Since the theoretical developments underpinning MIXSIM are only valid for Gaussian mixtures, we used it only to simulate bivariate Gaussian

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

Table 1. Performance of the bootstrap-based methodology, gap statistic with and without svd, and KL method for estimating K homogeneous spherical clusters for different settings. The left half of the table provides the mean ($\bar{\mathcal{K}}$) and the corresponding standard deviation ($\sigma_{\hat{\mathcal{K}}}$) of the estimated number of clusters, while the right half shows the mean ($\bar{\mathcal{D}}$) and standard deviation ($\sigma_{\mathcal{D}}$) of the adjusted Rand index deficiencies.

	$\breve{\omega}$	$\bar{\mathcal{K}}(\sigma_{\hat{\mathcal{K}}})$				$\bar{\mathcal{D}}(\sigma_{\mathcal{D}})$			
		$K=6, n=500$		$K=11, n=1000$		$K=6, n=500$		$K=11, n=1000$	
		$p=5$	$p=10$	$p=5$	$p=10$	$p=5$	$p=10$	$p=5$	$p=10$
Bootstrap	0.010	6.00 (0.00)	6.00 (0.00)	11.00 (0.00)	11.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	0.050	5.94 (0.58)	6.00 (0.00)	10.80 (0.45)	11.00 (0.00)	0.00 (0.03)	0.00 (0.00)	0.01 (0.02)	0.00 (0.00)
	0.100	5.96 (0.28)	5.97 (0.22)	10.81 (0.49)	10.71 (0.57)	0.01 (0.03)	0.01 (0.03)	0.01 (0.02)	0.02 (0.03)
Gap	0.010	5.99 (0.10)	6.00 (0.00)	10.96 (0.40)	11.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.02)	0.00 (0.00)
	0.050	5.91 (0.55)	6.00 (0.00)	10.77 (0.63)	11.00 (0.00)	0.00 (0.01)	0.00 (0.00)	0.01 (0.03)	0.00 (0.00)
	0.100	5.85 (0.52)	6.00 (0.00)	10.34 (1.67)	10.90 (1.00)	0.01 (0.04)	0.00 (0.00)	0.03 (0.12)	0.01 (0.09)
GapSVD	0.010	5.69 (1.20)	5.90 (0.70)	9.00 (4.02)	10.50 (2.19)	0.06 (0.23)	0.02 (0.14)	0.20 (0.40)	0.05 (0.22)
	0.050	5.77 (0.93)	5.75 (1.10)	10.16 (2.56)	10.20 (2.73)	0.02 (0.14)	0.05 (0.20)	0.07 (0.24)	0.08 (0.26)
	0.100	5.14 (1.67)	5.46 (1.55)	9.11 (3.55)	9.50 (3.59)	0.12 (0.27)	0.09 (0.27)	0.15 (0.31)	0.13 (0.31)
KL	0.010	5.99 (0.30)	6.02 (0.20)	10.89 (0.65)	11.05 (0.30)	0.01 (0.02)	0.00 (0.01)	0.01 (0.04)	0.00 (0.01)
	0.050	5.94 (0.75)	6.17 (0.53)	10.51 (1.48)	11.13 (0.44)	0.02 (0.07)	0.01 (0.03)	0.03 (0.07)	0.00 (0.01)
	0.100	5.68 (1.02)	6.46 (0.81)	10.01 (2.38)	11.21 (0.70)	0.05 (0.09)	0.03 (0.04)	0.07 (0.17)	0.01 (0.02)

mixtures with homogeneous spherical dispersions for $\breve{\omega} = 0.01, 0.05$ and 0.10. To generate datasets however, we replaced the Gaussian distribution in each coordinate with a t_ν distribution, where ν is the degrees of freedom. To facilitate illustration, we only considered $K = 5$, $p = 2$, $n = 100$. Our experimental suite consisted of the cases for which $\nu = 3, 10$ and ∞; the latter case is equivalent to realizations from a normal mixture. (Note that since MIXSIM is designed to simulate Gaussian mixtures, the actual overlap will be moderately to substantially higher, with decreasing ν, than the pre-specified levels of Gaussian overlap as the tails of the t_ν distribution are heavier than those of the normal distribution.) Nevertheless, performance evaluations on these datasets provide us with an opportunity to investigate the performance of the bootstrap-based procedure in assessing significance in the presence of heavy-tailed clusters vis-a-vis ν. Figure 4 provides k-means-clustered datasets to illustrate the level of complexity associated with t_ν-distributed clusters, for $\nu = 3$ and 10, and notional overlap $\breve{\omega} = 0.005$, 0.05, and 0.25. Figure 4 also provides the q-value quantitation maps (first columns) and the partitioning (second columns) at the corresponding bootstrap-significance-estimated $\hat{K}$ for each dataset. Clearly, the cases with $\nu = 3$ present substantially more complicated datasets to partition than those for $\nu = 10$. Note also that the quantitation maps in Figure 4 reflect clustering complexity very well. When $\breve{\omega}$ is smaller, we are more confident about the choice of K. The increase in ν also makes this choice easier. Thus, the two figures are also good illustrations of the use of quantitation maps in assessing significance in clustering.

The results of a more comprehensive simulation study over 100 datasets at each setting are summarized in Table 2 and in the supplement (Figure S-7). Clearly, the overall results agree with our expectations and with the initial impressions from Figure 4: the performance of our procedure is substantially better than that of all three competitors – *Gap*, *GapSVD* or KL. The improvement in performance of our approach becomes more pronounced with increased clustering complexity.

4.1.3 Performance with HDLSS datasets

Our next set of experiments in this setup of homogeneous spherical groups is a small-scale study for the case of clustering in high-dimensional datasets, specifically for when we have low sample sizes. For this set of experiments, we obtained 100 80-dimensional datasets that were simulated from 4-component Gaussian mixtures with $\breve{\omega} = 0.001$ and 0.01. Each dataset contained only 20 observations. The average number of detected clusters is 3.79 with $\sigma_{\hat{\mathcal{K}}} = 0.40$ for $\breve{\omega} = 0.001$ and 3.52 with $\breve{\omega} = 0.52$ for $\breve{\omega} = 0.01$ correspondingly. Further, $\bar{\mathcal{D}}$ was 0.08 with $\sigma_{\mathcal{D}} = 0.11$ for $\breve{\omega} = 0.01$ while $\bar{\mathcal{D}}$ was 0.03 with $\sigma_{\mathcal{D}} = 0.05$ for $\breve{\omega} = 0.001$. Thus, we have some indication that our procedure also works well in clustering datasets that fall within the large p, small n framework.

4.2 Nonhomogeneous compact clusters

We used MIXSIM to generate 100 simulated datasets at each setting of $(p, K, n, (\bar{\omega}, \breve{\omega}))$, where p, K and n were as before, while $(\bar{\omega}, \breve{\omega}) = (0.001, 0.005), (0.001, 0.01), (0.01, 0.05), (0.01, 0.1)$ and $(0.05, 0.25)$. For every simulated

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

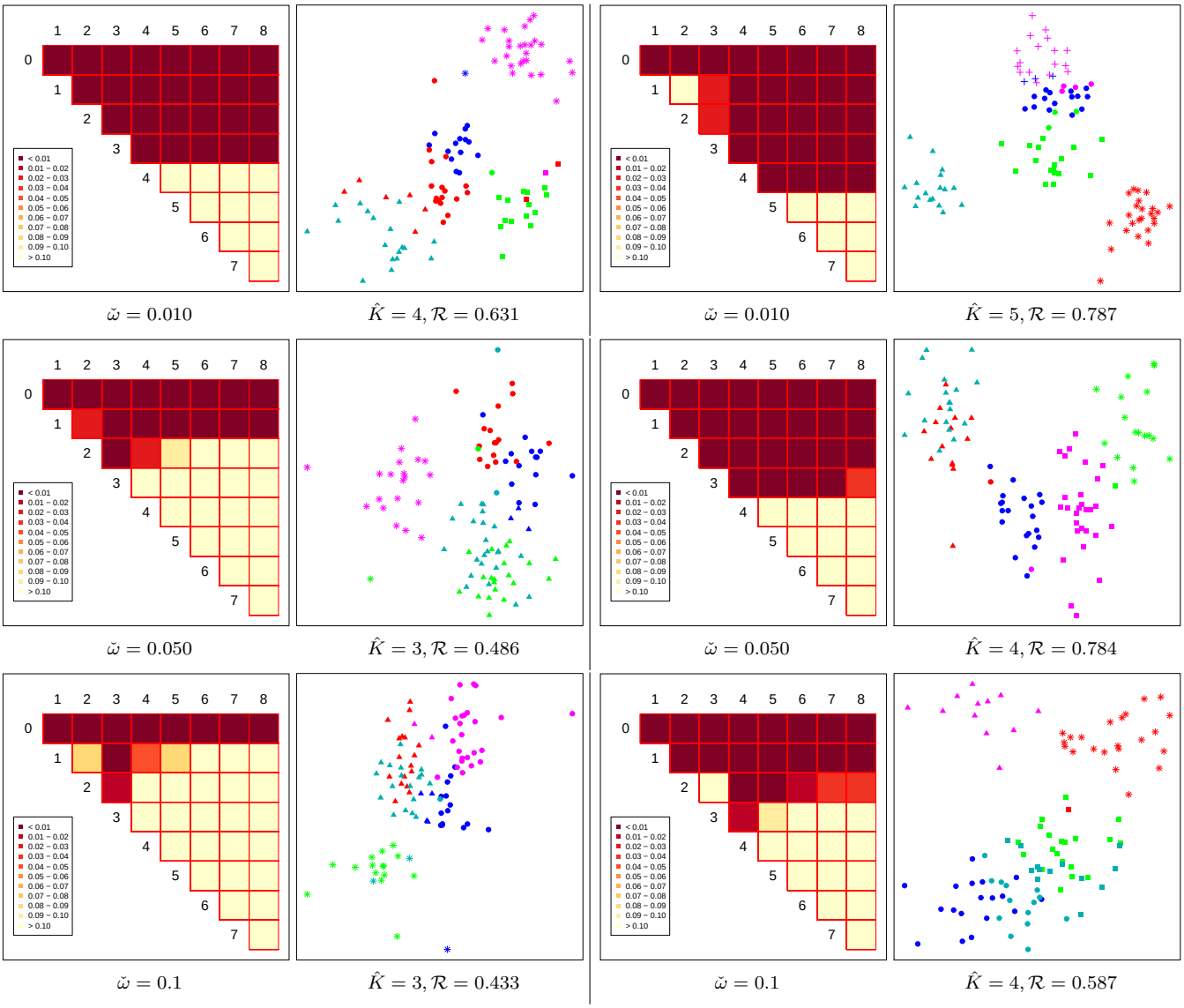


Figure 4. Quantitation maps and estimated classifications (for the optimal $\hat{K}$) for t-distributed datasets with 3 (left panel) and 10 (right panel) degrees of freedom with true $K = 5$, $p = 2$ and $n = 100$. In the figures representing the classification of the datasets, symbols represent true classification while colors illustrate estimated classification.

dataset, we used hierarchical clustering with Ward's criterion to partition the dataset into K groups for each K, followed by our bootstrap-based approach, *Gap*, and *GapSVD* to determine the suggested $\hat{K}$ by that method. For each of these datasets, we also used model-based clustering together with the Bayesian Information Criterion (BIC) to choose K. We denote this method by *BIC*. Table 3 provides numerical summaries of the performance of our bootstrap-based methodology, *Gap* and *GapSVD* and *BIC*, with further supporting evidence provided in the supplement (specifically, Figure S-7). For datasets with low $(\bar{\omega}, \breve{\omega})$, or high separation between clusters, we observe low values of $\mathcal{D}$ and good estimates for K. Expectedly, increased levels of overlap correspond to degraded performance all-around. Indeed, the proposed procedure underestimates K when overlap is substantial. That this is owing to clustering complexity and algorithm becomes clear when we note that $\mathcal{D}$-values remain remarkably low: this means that not all the clusters are always clearly distinguishable. The comparison of our bootstrap-based approach with *Gap*, *GapSVD* and *BIC* suggests that the bootstrap procedure performed better in many cases, and even when clusters are poorly-separated. This points to a clear preference for our bootstrap-based method over the gap statistic and BIC in complicated clustering cases. *Gap* and *GapSVD*. We remark again that as per Figure S-8, the latter often chooses $\hat{K} = 1$, especially when the overlap

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

Table 2. Performance of the bootstrap-based methodology, gap statistic with and without svd and KL method for estimating K with heavy-tailed clusters from the marginal t_ν distribution where ν represents the degrees of freedom of the t-distributions. Summaries are as in Table 1.

	$\tilde{\omega}$	$\bar{\mathcal{K}}(\sigma_{\hat{\mathcal{K}}})$			$\bar{\mathcal{D}}(\sigma_{\mathcal{D}})$		
		$\nu = 3$	$\nu = 10$	$\nu = \infty$	$\nu = 3$	$\nu = 10$	$\nu = \infty$
Bootstrap	0.010	4.77 (0.97)	4.86 (0.47)	4.90 (0.33)	0.03 (0.07)	0.01 (0.04)	0.01 (0.03)
	0.050	4.49 (1.17)	4.71 (0.70)	4.81 (0.49)	0.06 (0.10)	0.02 (0.06)	0.01 (0.04)
	0.100	4.15 (1.25)	4.54 (0.89)	4.76 (0.55)	0.09 (0.10)	0.04 (0.08)	0.02 (0.04)
Gap	0.010	3.72 (1.68)	4.08 (1.48)	4.31 (1.32)	0.23 (0.33)	0.18 (0.33)	0.13 (0.31)
	0.050	3.20 (1.56)	3.95 (1.44)	4.07 (1.37)	0.27 (0.30)	0.16 (0.28)	0.16 (0.30)
	0.100	2.95 (1.44)	3.78 (1.37)	3.91 (1.44)	0.28 (0.28)	0.16 (0.25)	0.17 (0.29)
GapSVD	0.010	3.47 (1.81)	3.66 (1.74)	4.22 (1.42)	0.30 (0.37)	0.28 (0.41)	0.16 (0.34)
	0.050	2.87 (1.57)	3.85 (1.59)	4.11 (1.48)	0.34 (0.34)	0.19 (0.33)	0.17 (0.34)
	0.100	2.69 (1.56)	3.69 (1.49)	3.87 (1.50)	0.35 (0.31)	0.19 (0.30)	0.18 (0.32)
KL	0.010	4.76 (1.30)	4.75 (1.08)	4.86 (0.68)	0.10 (0.15)	0.08 (0.14)	0.03 (0.07)
	0.050	4.21 (1.31)	4.55 (1.14)	4.53 (1.22)	0.13 (0.17)	0.09 (0.12)	0.12 (0.16)
	0.050	4.19 (1.35)	4.14 (1.33)	4.48 (1.33)	0.12 (0.13)	0.14 (0.16)	0.13 (0.15)

Table 3. Performance of the methodology in estimating the number of heterogeneous compact clusters. Summaries are as in Table 1.

	$\bar{\omega} : \tilde{\omega}$	$\bar{\mathcal{K}}(\sigma_{\hat{\mathcal{K}}})$				$\bar{\mathcal{D}}(\sigma_{\mathcal{D}})$			
		$K = 6, n = 500$		$K = 11, n = 1000$		$K = 6, n = 500$		$K = 11, n = 1000$	
		$p = 5$	$p = 10$	$p = 5$	$p = 10$	$p = 5$	$p = 10$	$p = 5$	$p = 10$
Bootstrap	0.001 : 0.005	6.00 (0.00)	6.00 (0.00)	10.96 (0.35)	11.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.03)	0.00 (0.00)
	0.001 : 0.01	5.99 (0.10)	6.00 (0.00)	11.01 (0.56)	11.00 (0.14)	0.00 (0.02)	0.00 (0.00)	0.01 (0.04)	0.00 (0.01)
	0.01 : 0.05	5.66 (0.74)	5.75 (0.54)	10.63 (1.36)	10.61 (0.99)	0.04 (0.05)	0.02 (0.03)	0.05 (0.05)	0.02 (0.04)
	0.01 : 0.1	5.69 (0.83)	5.68 (0.95)	10.24 (1.30)	10.70 (1.10)	0.04 (0.05)	0.04 (0.04)	0.03 (0.04)	0.03 (0.04)
	0.05 : 0.25	4.80 (1.31)	5.01 (1.71)	8.15 (2.29)	8.20 (2.16)	0.06 (0.06)	0.06 (0.07)	0.05 (0.03)	0.04 (0.04)
Gap	0.001 : 0.005	6.29 (0.54)	6.16 (0.37)	11.10 (0.98)	11.20 (0.40)	0.04 (0.07)	0.03 (0.06)	0.03 (0.08)	0.03 (0.06)
	0.001 : 0.01	6.25 (0.46)	6.26 (0.48)	11.23 (0.51)	11.22 (0.44)	0.03 (0.06)	0.03 (0.05)	0.02 (0.04)	0.03 (0.05)
	0.01 : 0.05	6.15 (0.61)	6.20 (0.45)	9.00 (2.33)	10.46 (0.99)	0.03 (0.06)	0.03 (0.05)	0.08 (0.15)	0.03 (0.04)
	0.01 : 0.1	6.38 (0.93)	6.74 (0.85)	9.11 (2.28)	10.62 (0.98)	0.04 (0.09)	0.03 (0.04)	0.08 (0.16)	0.03 (0.04)
	0.05 : 0.25	4.25 (1.60)	4.00 (1.92)	4.06 (2.03)	5.08 (2.14)	0.12 (0.18)	0.13 (0.18)	0.18 (0.16)	0.12 (0.12)
GapSVD	0.001 : 0.005	5.89 (1.31)	6.00 (0.55)	10.59 (2.64)	11.14 (0.35)	0.08 (0.24)	0.02 (0.11)	0.10 (0.24)	0.02 (0.06)
	0.001 : 0.01	6.14 (0.89)	6.10 (0.33)	10.61 (2.72)	11.00 (1.07)	0.05 (0.15)	0.02 (0.05)	0.10 (0.24)	0.03 (0.10)
	0.01 : 0.05	5.91 (1.15)	5.98 (0.79)	6.97 (3.96)	8.81 (3.42)	0.06 (0.16)	0.04 (0.12)	0.26 (0.32)	0.13 (0.26)
	0.01 : 0.1	5.99 (1.64)	6.25 (1.09)	8.11 (3.43)	9.61 (2.55)	0.10 (0.22)	0.04 (0.12)	0.17 (0.26)	0.08 (0.19)
	0.05 : 0.25	3.03 (1.80)	2.45 (1.62)	2.91 (2.07)	2.64 (1.94)	0.26 (0.25)	0.28 (0.21)	0.29 (0.18)	0.26 (0.15)
BIC	0.001 : 0.005	5.99 (0.30)	3.74 (0.56)	10.18 (1.02)	6.19 (0.68)	0.01 (0.07)	0.26 (0.10)	0.03 (0.04)	0.23 (0.06)
	0.001 : 0.01	6.09 (0.62)	3.96 (0.60)	10.37 (0.88)	6.10 (0.67)	0.02 (0.09)	0.24 (0.09)	0.03 (0.08)	0.24 (0.07)
	0.01 : 0.05	5.29 (0.83)	2.84 (0.44)	8.06 (1.13)	4.01 (0.64)	0.07 (0.11)	0.34 (0.10)	0.06 (0.06)	0.32 (0.09)
	0.01 : 0.1	5.27 (0.65)	3.47 (0.69)	7.90 (1.13)	4.27 (0.68)	0.05 (0.07)	0.21 (0.10)	0.06 (0.05)	0.29 (0.09)
	0.05 : 0.25	3.80 (0.62)	2.02 (0.35)	4.80 (0.99)	2.10 (0.44)	0.06 (0.07)	0.24 (0.11)	0.09 (0.07)	0.27 (0.07)

between groups is high.

4.3 Experiments with one or no clusters

A reviewer wondered whether our proposed procedure is too liberal with a propensity to overestimate the number of clusters. To investigate this possibility, we evaluated performance in 100 simulated datasets with no clustering and with one cluster present. The settings for n and p were the same as above, with $K_{\max} = 7$ for all cases. Our bootstrap-based approach had cent percent accuracy, identifying correctly, none or one cluster in all cases. Being unable to identify zero clusters, both *Gap* and *GapSVD* methods were consistent in suggesting $\hat{K} = 1$ when presented with datasets having completely ungrouped observations. However, when the true $K = 1$, *GapSVD* was accurate but *Gap* substantially overestimated K.

The results of our simulation studies suggest that the proposed procedure is a useful parametric-distribution-free tool that allows for assessing p-values without introducing strict model assumptions about the distribution of the data. The method compares very well, especially in relation to some common competitors, actually often outperforming them for the case of highly overlapping clusters. We next provide some theoretical underpinning to our methodology.

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

5. CONSISTENCY RESULTS

5.1 On the inconsistency of the naïve bootstrap

Here we show that the naïve residuals-based bootstrap typically fails to reproduce the null distribution of the test statistic. For concreteness, we work with $p = 2$ as in Figure S-1, and assume that there are K_0-many true spherical clusters, $\mathcal{G}_1, \ldots, \mathcal{G}_{K_0}$. For simplicity, we suppose that the points in the jth cluster are uniformly distributed over the disc $B(\boldsymbol{\mu}_j; r)$, centered at unknown $\boldsymbol{\mu}_j \in \mathbb{R}^2$ and that the clusters are well-separated and also that they all have the same scaling parameter.

Let $\widehat{\mathcal{G}}_j^{(K)}$, $i = 1, 2, \ldots, K$ denote the estimated partitions from a K-groups solution of the clustering algorithm. Denote the jth cluster center by $\hat{\boldsymbol{\mu}}_j$ and as before, let $\hat{\boldsymbol{\epsilon}}_i = \boldsymbol{X}_i - \sum_{k=1}^{K} \hat{\zeta}_{ik}^{(K)} \hat{\boldsymbol{\mu}}_k$, $i = 1, 2, \ldots, n$ be the residuals. In order to approximate the null distribution of the test statistic s_{K,K^*}, the null distribution of the resampled bootstrap variables must generate K spherical clusters. This, in particular, requires that the normalized residuals $\hat{\boldsymbol{\epsilon}}_i/\|\hat{\boldsymbol{\epsilon}}_i\|$, $i = 1, \ldots, n$ represent a random sample from the uniform distribution on the unit circle. Let $\hat{G}^{(K)}(\cdot)$ denote the empirical distribution of the normalized residuals $\hat{\boldsymbol{\epsilon}}_i/\|\hat{\boldsymbol{\epsilon}}_i\|$, $i = 1, 2, \ldots, n$ and let G denote the uniform distribution on the unit circle. The following theorem shows that for any $K < K^*$, and under some mild regularity conditions, the naïve bootstrap method fails to approximate G consistently.

Theorem 5.1. *Let $\mathcal{G}_j = \{\boldsymbol{X}_{ij} : i = 1, 2, \ldots, n_j\}$ denote the data in the jth cluster, so that $\boldsymbol{X}_{ij} = \boldsymbol{\mu}_j + \boldsymbol{\epsilon}_{ij}$ and $\{\boldsymbol{\epsilon}_{ij} : i = 1, 2, \ldots, n_j, j = 1, \ldots, K_0\}$ are independently and uniformly distributed on $B(\mathbf{0}; r)$. Let $n = n_1 + n_2 + \ldots + n_{K_0}$ and let $|B|$ denote the size of a finite set B. Suppose, further, that the following conditions hold for some $1 \leq K < K_0$:*

(C.1) There exists $\Delta > 0$ such that $B(\boldsymbol{\mu}_j, r + \Delta) \cap \mathcal{H}_{-j} = \emptyset$ where $\mathcal{H}_{-j}$ is the convex hull of $\{\boldsymbol{\mu}_k : k \neq j, 1 \leq k \leq K_0\}$.

(C.2) Suppose that for all $j = 1, 2, \ldots, K_0$, $n_j/n \to \pi_j^0 \in (0, 1)$ and that $\Delta > \left(\frac{\pi}{2(\pi_{\min}^0)^2} - 1\right) r$ where $\pi_{\min}^0 = \min\{\pi_j^0 : 1 \leq j \leq K_0\}$.

(C.3) Suppose that $\widehat{\mathcal{G}}_j^{(K)} \subset$ cone$(\theta_j(K), \phi_j(K))$ for some $0 \leq \theta_1^{(K)} < \phi_1^{(K)} < \theta_2^{(K)} < \phi_2^{(K)} < \ldots < \theta_K^{(K)} < \phi_K^{(K)} \leq 2\pi\}$ where cone(θ, ϕ) is the set of all points in $\mathbb{R}^2$ that lie in the cone emanating from the origin with an angle $\in [\theta, \phi]$ (with the x-axis).

Then, $\exists$ a constant $\delta_0 > 0$, depending only on $\{(\boldsymbol{\mu}_j, \pi_j^0) : j = 1, \ldots, K_0\}$, r, and K such that

$$P\Big(\liminf_{n \to \infty} \rho(\hat{G}^{(K)}, G) > \delta_0\Big) = 1,$$

where $\rho(\cdot)$ denotes the Prohorov metric (Billingsley, 1999) on the set of all probability measures on the unit circle and where π_j^0's are as in Condition (C.2).

Proof. See Appendix A. □

We discuss the implications of Theorem 5.1 and its conditions. Note that (C.1) above requires that the clusters be well-separated while (C.2) says that in the limit there are K_0 nontrivial clusters and that the parameter Δ is large compared to the radius r of each population cluster. Condition (C.3) stipulates that the clustering algorithm groups points from the adjacent clusters $\mathcal{G}_j$'s and that the resulting clusters each contain one or more complete clusters $\mathcal{G}_j$s. (We note that it is possible to prove a version of the theorem for clusters formed with split $\mathcal{G}_j$-s, but only under additional conditions on the structure of the clustering algorithm.) However, when the original clusters $\mathcal{G}_j$s are well-separated, Condition (C.3) holds for many clustering algorithms. In particular, it always holds for any clustering algorithm when $K = 1$ and for any given $K < K_0$ for the k-means algorithm, under suitable configurations of $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{K_0}$.

It also follows from Theorem 5.1 that the empirical distribution of the normalized residuals from the naïve bootstrap fails to approximate the uniform distribution on the unit circle even in the (weak) form of convergence in distribution, as the sample size becomes infinitely large. As a result, the naïve bootstrap method fails to give a valid approximation to the null distribution of the test statistic s_{K,K^*} when the conditions of the Theorem are satisfied, and therefore, any step-up procedure based on the naïve bootstrap method for quantitation is inconsistent. In comparison, the uniformity

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

of the normalized residuals is directly built into the formulation of the modified bootstrap method proposed in this paper and satisfies the key consistency condition:

$$\rho(\tilde{G}^{(K)}, G) \to 0 \quad as \;\; n \to \infty, \quad a.s. \tag{3}$$

for all $K < K_0$ where $\tilde{G}^{(K)}$ is the conditional distribution of the resampled error variable $\boldsymbol{\epsilon}_1^*$ under the modified bootstrap method. Indeed, as the modified bootstrap error variables have the same distribution as $\boldsymbol{Z}/\|\boldsymbol{Z}\|$, where $\boldsymbol{Z}$ has the standard multivariate normal distribution on $\mathbb{R}^d$, it follows that $\rho(\tilde{G}^{(K)}, G) = 0$ for every K and every n, so that (3) holds trivially.

5.2 Consistency of our suggested bootstrap procedure

Next we consider consistency of our suggested bootstrap procedure in somewhat more generality than in Section 5.1. Specifically, we suppose that under the null hypothesis, there are K_0 independent groups $\mathcal{G}_j = \{\boldsymbol{X}_{ij} : i = 1, \ldots, n_j\}$, $1 \le j \le K_0$ and $\boldsymbol{X}_{ij}$ has density $g(\boldsymbol{x} - \boldsymbol{\mu}_j)$ for all i, j, where $g(\cdot)$ is as in (1). Let $\hat{\mathcal{G}}_j$, $j = 1, 2, \ldots, K_0$ denote the estimated groups from the K_0-cluster solution, with respective cluster centers $\hat{\boldsymbol{\mu}}_j$'s. To approximate the *null* distribution of the test statistic $S_{K_0,K}$ by our modified bootstrap, we follow the steps described in Section 2.3. Specifically, we generate the bootstrap error variables $\{\boldsymbol{\epsilon}_1^*, \boldsymbol{\epsilon}_2^*, \ldots, \boldsymbol{\epsilon}_n^*\}$ as $\boldsymbol{\epsilon}_i^* = \|\hat{\boldsymbol{\epsilon}}_{\ell_i}\|\mathbf{W}_i$, $i = 1, 2, \ldots, n$, where the residuals $\{\hat{\boldsymbol{\epsilon}}_i : i = 1, 2, \ldots, n\}$ are obtained from the K_0-cluster solution of the clustering algorithm and where $\{\ell_1, \ell_2, \ldots, \ell_n\}$ is a (nonrandom) permutation of $\{1, 2, \ldots, n\}$. We next show that under some regularity conditions, the (conditional) distribution of the bootstrap error variable $\boldsymbol{\epsilon}_1^*$ under the modified bootstrap scheme provides a valid approximation to the null distribution of the error variable $\boldsymbol{\epsilon}_1$. To that end, for $j, k = 1, 2, \ldots, K_0$, let $\hat{\pi}_{j,k} = n^{-1}\sum_{i=1}^{n_j} \mathcal{I}(\boldsymbol{X}_i \in \hat{\mathcal{G}}_k)$ denote the proportion of $\boldsymbol{X}_i$'s from group $\mathcal{G}_j$ falling in the estimated cluster $\hat{\mathcal{G}}_k$. Also, let P_* denote the bootstrap probability and $\mathcal{G}$ denote the collection of all convex measurable subsets of $\mathbb{R}^p$. Then we have the following

Theorem 5.2. *Suppose that under H_0, $\mathcal{G}_j = \{\boldsymbol{X}_{ij} : i = 1, 2, \ldots, n_j\}$, $1 \le j \le K_0$ are independent and $\boldsymbol{X}_{ij}$ has density $g(\boldsymbol{x} - \boldsymbol{\mu}_j)$ for all i, j, where $g(\cdot)$ is as in (1) and where $n_j/n \to \pi_j^0 \in (0, 1)$ for all j. Further, assume the following conditions:*

(C.4) *(i) For all $j = 1, 2, \ldots, K_0$, $\hat{\boldsymbol{\mu}}_j \to \boldsymbol{\mu}_j$ almost surely.*

(ii) For all $j, k = 1, 2, \ldots, K_0$, $\exists\, \pi_{j,k} \in [0, 1]$ with $\pi_{j,j} = \pi_j^0 \ni \hat{\pi}_{j,k} \to \pi_{j,k}$ almost surely.

Then,

$$\sup_{C \in \mathcal{G}} \left| P_*(\boldsymbol{\epsilon}_1^* \in C) - P(\boldsymbol{\epsilon}_1 \in C) \right| \to 0 \quad as \quad n \to \infty, \quad \textit{almost surely.}$$

Proof. See Appendix B. □

Condition (C.4) is a condition on the clustering algorithm: Specifically, Part (i) of (C.4) says that under H_0, the cluster centers converge to the true cluster centers almost surely. (In particular, this condition holds for the k-means algorithm (cf. Pollard, 1981)). Additionally, Part (ii) of Condition (C.4) says that the clustering algorithm reproduces the proportion of points in the true clusters, asymptotically. Since $\sum_{k=1}^{K_0} \hat{\pi}_{j,k} = n_j/n \to \pi_j^0 = \pi_{j,j} \forall j$, it follows that $\sum_{k \ne j} \pi_{j,k} = 0$. As a result, under (C.4)(ii), the number of points in the jth cluster that are wrongly clustered is $o(n)$ for every $j = 1, 2, \ldots, K_0$. This is a weak requirement on the clustering algorithm which allows a large number of points to be clustered incorrectly, for large n.

Theorem 5.2 shows that under the conditions stated above, the modified bootstrap algorithm can successfully capture the distribution of the error variables, almost surely. Since the test statistic $s_{K_0,K}$ is a smooth function of the error variables $\{\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_n\}$ (given, in the case of this paper, by a difference of sums of squares), it follows that the modified bootstrap method can be used to approximate the null distribution of $s_{K_0,K}$. In contrast, as shown by Theorem 5.1. the naïve bootstrap method typically fails to reproduce the null distribution of the errors and hence, fails to provide a valid reference distribution for the test statistic $s_{K_0,K}$.

6. APPLICATION TO COLOR QUANTIZATION

Color quantization is used in computer graphics to reduce the number of colors in an image without appreciably losing its visual quality. The importance of this process comes from a need to display images on devices that are not

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

completely capable of dealing with multicolor images. It is also used for some image storing standards such as the Graphics Interchange Format (GIF).

Each pixel in an image is represented in terms of a mixture of red, green and blue colors with different intensities. This way of storing a color is known as *RGB* format. Hence, every picture can be presented as a three-dimensional dataset with the number of observations depending on the size of the picture. For example, an image of size 256×256 can be transformed into a dataset with 65,536 observations and a 512×512 image can be represented as a dataset of size 262,144. k-means color quantization applies the k-means algorithm, suitably initialized, to such a dataset to yield a palette of k colors for representing the image. Further, while k-means represents one of several approaches to color quantization (Emre Celebi, 2011), note also that K needs to be specified in k-means color quantization.

We apply our bootstrapping-for-significance procedure of Section 2.3.1 to several images from the USC-SIPI Image Database. Using our quantitation map provides us with two approaches to representing images with a certain number of colors. For instance, the conservative approach (of choosing the K for which we fail to reject the null hypothesis against the alternative of $K + 1$ for the first time) provides us with the largest number of colors which represents the image significantly better than lesser number of colors. In some sense therefore, we may regard this K as providing the "minimal palette" or the minimum number of colors needed to display the image. Our preferred approach, applied with a $K_{max} = 100$, on the other hand provides us with the fewest number of colors in the palette above which there is not much significant improvement in image quality. This may be regarded, in the same spirit as the minimal palette, as providing the "optimal palette" for the image. Figure 5 provides results for six images in the

Figure 5. Color quantization results for Tree, Couple, House, Lena, Baboon and Peppers images. The first row represents original images, the second row provides images obtained using a minimal palette of K colors ($K = 6, 7, 6, 8, 7, 9$ respectively), while the last row displays images using our optimal palette, with K colors ($K = 26, 39, 27, 21, 31, 17$, respectively).

database: these are *Tree*, *Couple*, *House* (each on a grid of 256×256 pixels) and *Lena*, *Baboon* and *Peppers* (each on a grid of 512×512 pixels). The first row presents the original images, while the second row provides images obtained with pixel values replaced by the colors in our minimal palette, which consisted of $K = 6$ for *Tree*, $K = 7$ for *Couple*, $K = 6$ for *House*, $K = 8$ for *Lena*, $K = 7$ for *Baboon* and $K = 9$ for *Peppers*. The third row provides images using our optimal palette which chose $K = 26$ colors for *Tree*, $K = 39$ for *Couple*, $K = 27$ for *House*, $K = 21$ for *Lena*, $K = 31$ for *Baboon*, and $K = 17$ for *Peppers*. As we can see, the images from the middle row are reasonable but not of excellent quality. The images in the third row represented using our optimal palette are each of much better quality, and mostly visually very close to the originals. Overall, the performance of our procedure produced pictures of very

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

reasonable quality given the number of colors involved.

7. CONCLUSION

In this paper, we develop methodology for assessing significance in compact clusters through a nonparametric bootstrap procedure. The basic strategy compares any two models in a testing framework and recommends the more complicated model only if we observe a significant p-value. The naïve bootstrap approach for this problem has the drawback of having very low power, so we develop an approach that exploits the compactness of groups inherent in a clustering model. We first develop methodology for the case of spherical homogeneous clusters and then extend it to the more general case of nonhomogeneous ellipsoidal groups. We also develop *quantitation maps* based on the p- and q-values which can provide researchers with a comprehensive display of the relative strengths of a complicated model vis-a-vis a simpler one. It can also be used to estimate the total number of groups in a dataset. Our methodology was illustrated on two classification datasets and evaluated very thoroughly in a series of simulation experiments. For comparative purposes, we evaluated performance of our methodology in terms of its ability to estimate the number of clusters in the dataset. The proposed approach was seen to edge out its competitors quite often: the improvement was very emphatic even when the clustering complexity of the dataset was high. Further, we developed theoretical results to show that our developed bootstrap methodology was consistent. Finally, we also applied our methodology to determine the minimum and optimal number of colors in a palette to represent RGB images, with excellent results.

There are several additional areas in which we could use our methodology. For instance we could use the development to study the importance of each coordinate in clustering a dataset. We could also modify our approach to assess significance in the case of semi-supervised clustering where some of the class information has been observed in the labeled part of the dataset, but it is not known if, for instance, there are classes that have not yet been observed at all. Another issue would be to investigate, as pointed out by a reviewer, how to extend the methodology to the case for general non-Euclidean non-Mahalanobis distance clustering, *eg*, where the observations are discrete. Such generalization may be possible in certain cases. For instance, it may be possible to apply our methodology in the context of certain versions of spectral clustering (von Luxburg, 2007), where the problem reduces to k-means clustering of the first k-eigenvectors of the similarity matrix. In other scenarios, our methodology may need to be substantially developed and extended further. In any case, this is another interesting area for further investigation.

There are some other areas that are in need of further study. For instance, another reviewer has asked about the fate of our methodology for the general (ellipsoidal) case with HDLSS data. In such cases, of course, the group-specific dispersion matrices can not be inverted. However, it is our view that clustering based on formal procedures in this setting is meaningful only within the framework of additional assumptions (*eg*, a lower-dimensional representation for the dispersions) and those assumptions can then make it possible to obtain an alternative representation of $\mathbf{\Sigma}^{-1}$. Our methodology should then be possible to apply using these modifications. Of course, it would be important to explore this aspect further. Of interest also would be to explore the case when clusters are more general than ellipsoidal, as commented on by a third reviewer. In Section S-5, we presented an example where it was possible to find an appropriate Ψ (via the multivariate Box-Cox transform) and where our methodology provided very good results. It would be important to investigate and see if this performance is sustained in more cases. Thus, we see that while our paper has made a significant contribution to developing and using the bootstrap for assessing significance of compact clusterings, a few interesting issues worthy of further attention remain.

APPENDIX A: PROOF OF THEOREM 5.1

Since $K < K_0$, by Condition (C.3), $\exists$ a cluster $\widehat{\mathcal{G}}$ (in $\{\widehat{\mathcal{G}}_1, \widehat{\mathcal{G}}_2, \ldots, \widehat{\mathcal{G}}_K\}$) that contains at least two of the clusters $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_{K_0}$. Fix such a $\widehat{\mathcal{G}}$ and without loss of generality (w.l.g.), suppose that $\widehat{\mathcal{G}} = \cup_{j=1}^{s} \mathcal{G}_j$ for some $2 \leq s \leq K_0$. Then the residuals from the cluster $\widehat{\mathcal{G}}$ are given by

$$\hat{\boldsymbol{\epsilon}}_{ij} = \boldsymbol{X}_{ij} - \hat{\boldsymbol{\mu}} = \boldsymbol{\epsilon}_{ij} + [\boldsymbol{\mu}_j - \hat{\boldsymbol{\mu}}],$$

where $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, s$. By the Laws of Large Numbers and Condition (C.2), we have

$$\hat{\boldsymbol{\mu}} = \sum_{j=1}^{s}\sum_{i=1}^{n_j} X_{ij} / \sum_{j=1}^{s} n_j = \sum_{j=1}^{s}\sum_{i=1}^{n_j} \boldsymbol{\epsilon}_{ij} / \sum_{j=1}^{s} n_j + \sum_{j=1}^{s} n_j \boldsymbol{\mu}_j / \sum_{j=1}^{s} n_j = \sum_{j=1}^{s} p_j \boldsymbol{\mu}_j + o(1) \quad \text{a.s.}$$

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

where $p_j \equiv \pi_j^0 / \sum_{k=1}^s \pi_k^0 \in (0,1)$, $j = 1, 2, \ldots, s$.

Next using (A.1) and Condition (C.2), it can be shown that there exists a sequence $t_n \downarrow 0$ as $n \to \infty$ such that for any Borel set A of the unit circle

$$\begin{aligned}
\hat{G}(A) &\geq n^{-1} \sum_{j=1}^{s} \sum_{i=1}^{n_j} \mathcal{I}(\hat{\boldsymbol{\epsilon}}_{ij}/\|\hat{\boldsymbol{\epsilon}}_{ij}\| \in A) \\
&\geq n^{-1} n_1 P(\hat{\boldsymbol{\epsilon}}_{11}/\|\hat{\boldsymbol{\epsilon}}_{11}\| \in A) + o(1) \quad \text{a.s.} \\
&\geq \pi_1^0 \cdot P([\boldsymbol{\epsilon}_{11} + \boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}]/\|\boldsymbol{\epsilon}_{11} + \boldsymbol{\mu}_1 - \hat{\boldsymbol{\mu}}\| \in A) + o(1) \quad \text{a.s.,} \\
&\geq \pi_1^0 \cdot P(\boldsymbol{U}/\|\boldsymbol{U}\| \in A^{-t_n}) + o(1) \quad \text{a.s.,} \qquad \text{(A.1)}
\end{aligned}$$

where $A^{-t} = \{\boldsymbol{x} \in A : B(\boldsymbol{x}; t) \subset A\}$, $t > 0$, and where $\boldsymbol{U}$ has the uniform distribution on $B(\boldsymbol{a}; r)$ with $\boldsymbol{a} = \boldsymbol{\mu} - \sum_{j=1}^s p_j \boldsymbol{\mu}_j$. W.l.g, suppose that $\boldsymbol{a} = (a_1, a_2)' \in (0, \infty)^2$. Then, using Conditions (C.1) and (C.2), it can be shown that $B(\boldsymbol{a}; r) \subset (0, \infty)^2$ and that $\|\boldsymbol{a}\| > r$. Further, it is not difficult to check that the distribution of $\boldsymbol{U}/\|\boldsymbol{U}\|$ has a density, given by

$$f(\theta) = \frac{2A(\theta)B(\theta)}{\pi r^2 [1 + m(\theta)^2]} \mathcal{I}(|\theta - \theta_0| \leq \theta_1)$$

where $\theta_0 = \tan^{-1}(a_2/a_1)$, $\theta_1 = \sin^{-1}(r/\|\boldsymbol{a}\|)$, $m(\theta) = \tan\theta$, and $A(\theta) = a_1 + m(\theta)a_2$ and $B(\theta) = (A(\theta_0)^2 - [1 + m(\theta)^2](\|\boldsymbol{a}\|^2 - r^2))^{\frac{1}{2}}$. Using Condition (C.2), choose a $\rho \in (0,1)$ such that

$$\rho(\Delta + r) > \pi r / [2(\pi_{\min}^0)^2]. \qquad \text{(A.2)}$$

Since $f(\theta_0) = \frac{2\|\boldsymbol{a}\|}{\pi r}$, there exists a $\eta > 0$ (depending only on $\boldsymbol{a}$ r and ρ) such that $f(\theta) > \frac{2\rho\|\boldsymbol{a}\|}{\pi r}$ for all $|\theta - \theta_0| \leq \eta$. Then, from (A.1), it follows that with $A = (\theta_0 - \eta, \theta_0 + \eta)$,

$$\begin{aligned}
\hat{G}(A) &\geq \pi_1^0 \cdot P(\boldsymbol{U}/\|\boldsymbol{U}\| \in A^{-t_n}) + o(1) \quad \text{a.s.,} \\
&\geq \pi_1^0 \cdot \frac{2\rho\|\boldsymbol{a}\|}{\pi r} \cdot (2\eta) + o(1) \quad \text{a.s.,} \\
&\geq G(A) + \delta_1 + + o(1) \quad \text{a.s.,} \qquad \text{(A.3)}
\end{aligned}$$

where $\delta_1 = 2\eta\Big(\frac{2\pi_1^0 \rho\|\boldsymbol{a}\|}{\pi r} - 1\Big)$.

Next, note that by Condition (C.1),

$$\begin{aligned}
\|\sum_{j=1}^{s} p_j \boldsymbol{\mu}_j - \boldsymbol{\mu}_1\| &= \|(1 - p_1)\boldsymbol{\mu}_1 - \sum_{j=2}^{s} p_j \boldsymbol{\mu}_j\| \\
&= (1 - p_1)\|\boldsymbol{\mu}_1 + \sum_{j=2}^{s} (1 - p_1)^{-1} p_j \boldsymbol{\mu}_j\| \\
&\geq (1 - p_1) \cdot \inf\{\|\boldsymbol{\mu}_1 - \boldsymbol{x}\| : \boldsymbol{x} \in \mathcal{H}_{-1}\} \\
&\geq (1 - p_1)[r + \Delta]. \qquad \text{(A.4)}
\end{aligned}$$

Hence, by (A.2) and Condition (C.2),

$$\begin{aligned}
\delta_1 \cdot \frac{\pi r}{2\eta} &= 2\pi_1^0 \rho\|\boldsymbol{a}\| - \pi r \\
&\geq 2\pi_1^0 \rho(1 - p_1)[r + \Delta] - \pi r \\
&> 2\pi_1^0 (1 - p_1) \frac{\pi r}{[2(\pi_{\min}^0)^2]} - \pi r \\
&= \pi r \Big[\frac{\pi_1^0 (1 - p_1)}{(\pi_{\min}^0)^2} - 1\Big] \geq 0,
\end{aligned}$$

as $\pi_1^0(1 - p_1) = \pi_1^0 \sum_{j=2}^s \pi_j^0 / \sum_{j=1}^s \pi_j^0 \geq \pi_1^0 \pi_2^0 / 1 \geq (\pi_{\min}^0)^2$. Hence, $\delta_1 > 0$ and Theorem 5.1 follows from (A.3) by taking $\delta_0 \in (0, \delta_1)$.

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

APPENDIX B: PROOF OF THEOREM 5.2

Let τ denote the inverse permutation to $\{\ell_1, \ell_2, \ldots, \ell_n\}$, i.e., $\tau(\ell_i) = i$ for $i = 1, 2, \ldots, n$. Let $\mathcal{I}(\cdot)$ denote the indicator function. Also, for any set $C \subset \mathbb{R}^p$ and $t \in (0, \infty)$, let $C^t = \{\boldsymbol{x} \in \mathbb{R}^p : \|\boldsymbol{x} - \mathbf{y}\| \le t \text{ for some } \mathbf{y} \in C\}$ denote the t-enlargement of C, and similarly, define $C^{-t} = \{\boldsymbol{x} \in C : B(\boldsymbol{x}; t) \subset C\}$ where $B(\boldsymbol{x}; t)$ is the closed ball of radius t centered at $\boldsymbol{x}$.

Note that by construction,

$$\begin{aligned}
P_*(\boldsymbol{\epsilon}_1^* \in C) &= n^{-1} \sum_{i=1}^{n} \mathcal{I}(\|\hat{\boldsymbol{\epsilon}}_{\ell_i}\| \mathbf{W}_i \in C) \\
&= n^{-1} \sum_{i=1}^{n} \mathcal{I}(\|\hat{\boldsymbol{\epsilon}}_i\| \mathbf{W}_{\tau(i)} \in C) \\
&= n^{-1} \sum_{j=1}^{K_0} \sum_{k=1}^{K_0} \sum_{i=1}^{n} \mathcal{I}(\boldsymbol{X}_i \in \mathcal{G}_j) \mathcal{I}(\boldsymbol{X}_i \in \hat{\mathcal{G}}_k) \mathcal{I}(\|\boldsymbol{\epsilon}_i + \boldsymbol{\mu}_j - \hat{\boldsymbol{\mu}}_k\| \mathbf{W}_{\tau(i)} \in C) \\
&= n^{-1} \sum_{j=1}^{K_0} \sum_{i=1}^{n} \mathcal{I}(\boldsymbol{X}_i \in \mathcal{G}_j) \mathcal{I}(\|\boldsymbol{\epsilon}_i + \boldsymbol{\mu}_j - \hat{\boldsymbol{\mu}}_j\| \mathbf{W}_{\tau(i)} \in C) + R_{1n}(C), \quad \text{(say)} \\
&\equiv \hat{F}_n(C) + R_{1n}(C), \quad \text{say,}
\end{aligned} \tag{A.5}$$

where $R_{1n}(C)$ is defined by subtraction and admits the bound

$$\begin{aligned}
\sup_{C \in \mathcal{G}} |R_{1n}(C)| &\le n^{-1} \sum_{1 \le j \ne k \le K_0} |\hat{\mathcal{G}}_j \cap \mathcal{G}_k| + n^{-1} \sum_{j=1}^{K_0} |\hat{\mathcal{G}}_j \Delta \mathcal{G}_j| \\
&\le \sum_{1 \le j \ne k \le K_0} \hat{\pi}_{j,k} + \sum_{j=1}^{K_0} |\hat{\pi}_{j,j} - \pi_{j,j}| \\
&= o(1) \quad \text{as} \quad n \to \infty, \quad \text{almost surely,}
\end{aligned}$$

by Condition (C.4)(ii).

Next fix $\delta \in (0, \infty)$ and define the event $A_n = \{\|\hat{\boldsymbol{\mu}}_j - \boldsymbol{\mu}_j\| \le \delta\}$. Also, for notational consistency, for $j = 1, 2, \ldots, K_0$, denote the set of $\mathbf{W}_{\tau(i)}$s corresponding to the indices i from the jth cluster (i.e., for all i with $\boldsymbol{X}_i \in \mathcal{G}_j$) by $\{\tilde{\mathbf{W}}_{ij} : i = 1, 2, \ldots, n_j\}$. Then, by Condition (C.4)(i), it follows that

$$P(A_n \text{ infinite often }) = 0. \tag{A.6}$$

Further, on A_n,

$$n^{-1} \sum_{j=1}^{K_0} \sum_{i=1}^{n_j} \mathcal{I}(\|\boldsymbol{\epsilon}_{ij}\| \tilde{\mathbf{W}}_{ij} \in C^{-\delta}) \le \hat{F}_n(C) \le n^{-1} \sum_{j=1}^{K_0} \sum_{i=1}^{n_j} \mathcal{I}(\|\boldsymbol{\epsilon}_{ij}\| \tilde{\mathbf{W}}_{ij} \in C^{\delta}). \tag{A.7}$$

Since τ is a nonrandom permutation, it follows that $\|\boldsymbol{\epsilon}_{ij}\| \tilde{\mathbf{W}}_{ij}$ are iid, with the same distribution as that of $\boldsymbol{\epsilon}_1$. Hence, by the (generalized) Glivenko-Cantelli theorem (cf. Elker et al., 1979; van der Vaart and Wellner, 1996),

$$\sup_{C \in \mathcal{G}} \left| n^{-1} \sum_{j=1}^{K_0} \sum_{i=1}^{n_j} \mathcal{I}(\|\boldsymbol{\epsilon}_{ij}\| \tilde{\mathbf{W}}_{ij} \in C) - P(\boldsymbol{\epsilon}_1 \in C) \right| = o(1) \quad \text{as} \quad n \to \infty, \quad \text{a.s.} \tag{A.8}$$

Since $\{C^{\pm\delta} : C \in \mathcal{G}, \delta > 0\} = \mathcal{G}$ and $\sup\{P(\boldsymbol{\epsilon}_1 \in C^{\delta} \setminus C^{-\delta}) : C \in \mathcal{G}\} = o(1)$ as $\delta \downarrow 0$ (cf. Bhattacharya and Rao, 2010), the theorem now follows from (A.5), (A.6), (A.7) and (A.8).

. References

Anderson, E. (1935), "The Irises of the Gaspe Peninsula," *Bulletin of the American Iris Society*, 59, 2–5.

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

Andrews, D. F. (1972), "Plots of High-dimensional Data," *Biometrics*, 28, 125–136.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, 57, 289–300.

Bhattacharya, R. N. and Rao, R. R. (2010), *Normal approximation and asymptotic expansions*, Philadelphia, PA: SIAM.

Biernacki, C., Celeux, G., and Govaert, G. (2003), "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models," *Computational Statistics and Data Analysis*, 413, 561–575.

Billingsley, P. (1999), *Convergence of Probability Measures*, New York: John Wiley & Sons, Inc.

Cambanis, S., Huang, S., and Simons, G. (1981), "On the theory of elliptically contoured distributions," *Journal of Multivariate Analysis*, 11, 368–385.

Cramer, H. (1946), *Mathematical methods of statistics*, Princeton, New Jersey: Princeton University Press.

Dudoit, S. and Fridlyand, J. (2003), "Bagging to improve the accuracy of a clustering procedure," *Bioinformatics*, 19, 1090–9.

Elker, J., Pollard, D., and Stute, W. (1979), "Glivenko-Cantelli theorems for classes of convex sets," *Advances in Applied Probability*, 11, 820–833.

Emre Celebi, M. (2011), "Improving the performance of k-means for color quantization," *Image and Vision Computing*, 29, 260–271.

Everitt, B. S. (1979), "Unresolved problems in cluster analysis," *Biometrics*, 35, 169–181.

Everitt, B. S., Landau, S., and Leesem, M. (2001), *Cluster Analysis (4th ed.)*, New York: Hodder Arnold.

Fisher, R. A. (1936), "The use of multiple measurements in taxonomic poblems," *Annals of Eugenics*, 7, 179–188.

Forgy, E. (1965), "Cluster analysis of multivariate data: efficiency vs. interpretability of classifications," *Biometrics*, 21, 768–780.

Fraley, C. and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631.

Haykin, S. (1999), *Neural networks: A comprehensive foundation*, Saddle River, NJ: Prentice Hall, 2nd ed.

Hinneburg, A. and Keim, D. (1999), "Cluster discovery methods for large databases: from the past to the future," in *Proceedings of the ACM SIGMOD International Conference on the Management of Data*.

Hubert, L. and Arabie, P. (1985), "Comparing partitions," *Journal of Classification*, 2, 193–218.

Jain, A. and Dubes, R. (1988), *Algorithms for clustering data*, Englewood Cliffs, NJ: Prentice Hall.

Johnson, S. (1967), "Hierarchical clustering schemes," *Psychometrika*, 32:3, 241–254.

Kaufman, L. and Rousseuw, P. J. (1990), *Finding Groups in Data*, New York: John Wiley and Sons, Inc.

Kerr, M. K. and Churchill, G. A. (2001), "Bootstrapping cluster analysis: Assessing reliability of conclusions from microarray experiments," *Proceedings of the National Academy of Sciences*, 98, 8961–8965.

Kettenring, J. R. (2006), "The practice of cluster analysis," *Journal of classification*, 23, 3–30.

Krzanowski, W. J. and Lai, Y. T. (1985), "A criterion for determining the number of groups in a data set using sum of squares clustering," *Biometrics*, 44, 23–34.

Liu, Y., Hayes, D. N., Nobel, A., and Marron, J. S. (2008), "Statistical Significance of Clustering for High-Dimensional, Low Sample Size Data," *Journal of the American Statistical Association*, 103, 1281–1293.

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

MacQueen, J. (1967), "Some methods for classification and analysis of multivariate observations," *Proceedings of the Fifth Berkeley Symposium*, 1, 281–297.

Maitra, R. (2001), "Clustering massive datasets with applications to software metrics and tomography," *Technometrics*, 43, 336–346.

— (2009), "Initializing Partition-Optimization Algorithms," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6, 144–157.

Maitra, R. and Melnykov, V. (2010a), "Assessing significance in finite mixture models," Tech. rep., Department of Statistics, Iowa State University.

— (2010b), "Simulating data to study performance of finite mixture modeling and clustering algorithms," *Journal of Computational and Graphical Statistics*, 19, 354–376.

Marriott, F. H. (1971), "Practical problems in a method of cluster analysis," *Biometrics*, 27, 501–514.

McLachlan, G. (1987), "On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture," *Applied Statistics*, 36, 318–324.

McLachlan, G. and Peel, D. (2000), *Finite Mixture Models*, New York: John Wiley and Sons, Inc.

McLachlan, G. J. and Basford, K. E. (1988), *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

McShane, L. M., Radmacher, M. D., Freidlin, B., Yu, R., Li, M.-C., and Simon, R. (2002), "Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data," *Bioinformatics*, 18, 1462–1469.

Michener, C. D. and Sokal, R. R. (1957), "A quantitative approach to a problem in classification," *Evolution*, 11, 130–162.

Milligan, G. W. and Cooper, M. C. (1985), "An examination of procedures for etermining the number of clusters in a dataset." *Psychometrika*, 50, 159–179.

Murtagh, F. (1985), *Multi-dimensional clustering algorithms*, Berlin; New York: Springer-Verlag.

Pollard, D. (1981), "Strong Consistency of K-Means Clustering," *Annals of Statistics*, 9, 135–140.

Ramey, D. B. (1985), "Nonparametric clustering techniques," in *Encyclopedia of Statistical Science*, New York: Wiley, vol. 6, pp. 318–319.

Rumelhart, D. and Zipser, D. (1985), "Feature discovery by competitive learning," *Cognitive Science*, 9, 75–112.

Ruspini, E. (1970), "Numerical methods for fuzzy clustering," *Information Science*, 2, 319–350.

Struyf, A., Hubert, M., and Rousseeuw, R. (1997), "Clustering in an Object-Oriented Environment," *Journal of Statistical Software*, 1.

Tibshirani, R. J. and Walther, G. (2005), "Cluster validation by prediction strength," *Journal of Computational and Graphical Statistics*, 14, 511–528.

Tibshirani, R. J., Walther, G., and Hastie, T. J. (2003), "Estimating the number of clusters in a dataset via the gap statistic," *Journal of the Royal Statistical Society*, 63, 411–423.

Titterington, D., Smith, A., and Makov, U. (1985), *Statistical Analysis of Finite Mixture Distributions*, Chichester, U.K.: John Wiley & Sons.

van der Vaart, A. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York, NY: Springer.

von Luxburg, U. (2007), "A Tutorial on Spectral Clustering," *Statistics and Computing*, 17, 395–416.

Xu, R. and Wunsch, D. C. (2009), *Clustering*, NJ, Hoboken: John Wiley and Sons, Inc.

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

Supplement to "Bootstrapping for Significance of Compact Clusters in Multi-dimensional Datasets"

Ranjan Maitra, Volodymyr Melnykov and Soumendra N. Lahiri *

S-1. HOMOGENEOUS SPHERICAL CLUSTERS

We provide several illustrations for the case of homogeneous spherical clusters obtained using the k-means algorithm. Figure S-1 represents sets of four replications, each set obtained under the null distribution with $H_0 : K = 1, 2, 3, 4$, respectively and by resampling with replacement *Ruspini* (Ruspini, 1970) dataset using the naïve approach described in Section 2.3.1 of the manuscript. Figures S-1a-d display sample replications under the null distribution assuming only one cluster, but we note that there are clearly four distinct clusters in all cases. Therefore, the naïve approach cannot be relied upon for simulating a meaningful null distribution. In particular, the test $H_0 : K = 1$ versus $H_a : K = 4$ will have low power and the null hypothesis is not likely to be rejected, even with a carefully chosen test statistic. A similar situation is observed for $H_0 : K = 2$. From Figure S-1, we can see that reference samples for $H_0 : K = 1$, $H_0 : K = 2$, and $H_0 : K = 4$ are similar. Things improve somewhat for $H_0 : K = 3$, in that the reference samples are much less similar than the original dataset. On the contrary, Figure S-2 provides datasets resampled from *Ruspini* dataset based on the proposed methodology for homogeneous spherical clusters employing k-means algorithm. The first row represents datasets simulated under the null hypothesis $H_0 : K = 1$, the second, third and forth rows assume $H_0 : K = 2$, $H_0 : K = 3$, and $H_0 : K = 4$ correspondingly. It can be clearly seen, that the null distribution for $K = 1$ (first row of Figure S-2) makes substantially more sense compared with the one illustrated in Figure S-1. There is one large cluster in all four cases. The lack of observations in the middle of the cluster follows from the fact that the geometrical center of the four true clusters is located in the area with no neighboring observations. The second and third rows of Figure S-2 also provide reasonable null distributions for $K = 2$ and $K = 3$ respectively. Meanwhile, they are very different in appearance from the original *Ruspini* dataset. At the same time, the datasets from the last row of plots are very much alike the original dataset they were resampled from. It explains why any null hypothesis with $H_0 : K = 1, 2, 3$ has to be rejected in favor of the alternative hypothesis $H_a : K = 4$.

S-2. GENERAL ELLIPSOIDAL CLUSTERS

In this section, we illustrate the proposed methodology on a simulated dataset presented at Figure S-3. The dataset consists of four clusters with 100 bivariate observations. Four plots provide true and estimated groupings obtained by model-based clustering for $K = 2, 3, 4$, and 5. True groupings are illustrated by color while estimated clustering is provided by different characters. As we can see, the clusters are fairly distinct but are far from being homogeneous spherical clusters. Figure S-4 illustrates datasets resampled from the original dataset using our methodology for general ellipsoidal clusters. Similarly to Figure S-2 illustrating an example with *Ruspini* dataset, we can see that the null distributions under $K = 1, 2$, and 3 (see the first three rows of Figure S-4 correspondingly) are all meaningful. However, only the last row of plots provides datasets that are very similar to the original dataset. Indeed, this explains why any of the null hypothesis $H_0 : K = 1, 2, 3$ has to be rejected in favor of the alternative $H_a : K = 4$.

S-3. ADDITIONAL DETAILS ON ILLUSTRATION USING IRIS DATA

S-3.1 A note on Andrews' curves

There are many ways of representing multivariate data (Wegman and Carr, 1993; Wegman et al., 1993; Theus, 2008). However they are all either heuristic or reduce the data into two or three dimensions and then display these

*Ranjan Maitra is Professor in the Department of Statistics and Statistical Laboratory, Iowa State University, Ames, IA 50011-1210, Volodymyr Melnykov is Assistant Professor in the Department of Statistics, North Dakota State University, Fargo, ND 50373, and Soumendra N. Lahiri is Professor in the Department of Statistics at Texas A& M University. This research was supported in part by the National Science Foundation CAREER Grant # DMS-0437555. and the National Institutes of Health Grant # DC-006740.

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

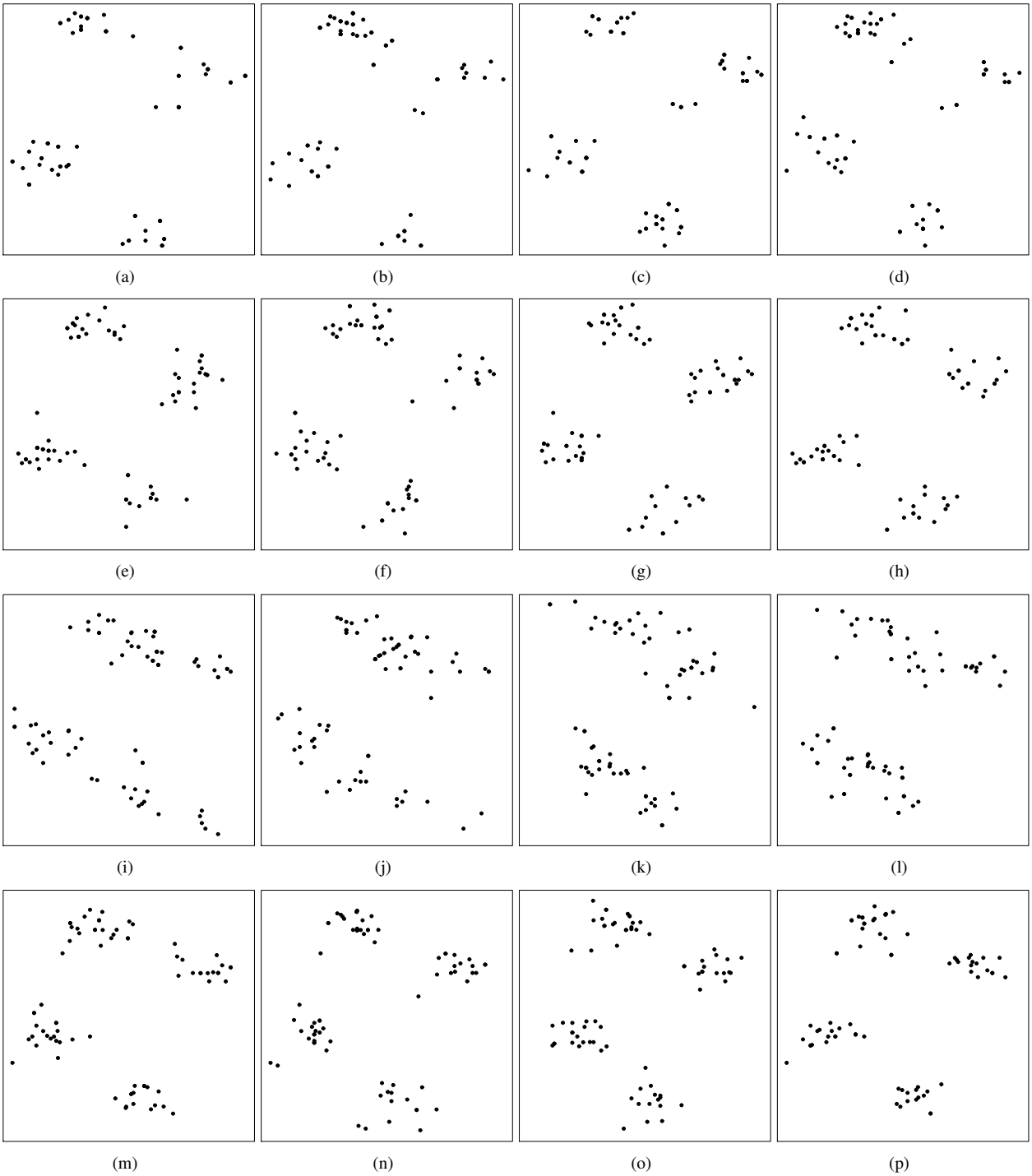


Figure S-1. Datasets simulated with the naïve approach for spherical clusters under the null hypotheses $H_0 : K = 1$ (first row), $H_0 : K = 2$ (second row), $H_0 : K = 3$ (third row), and $H_0 : K = 4$ (last row).

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

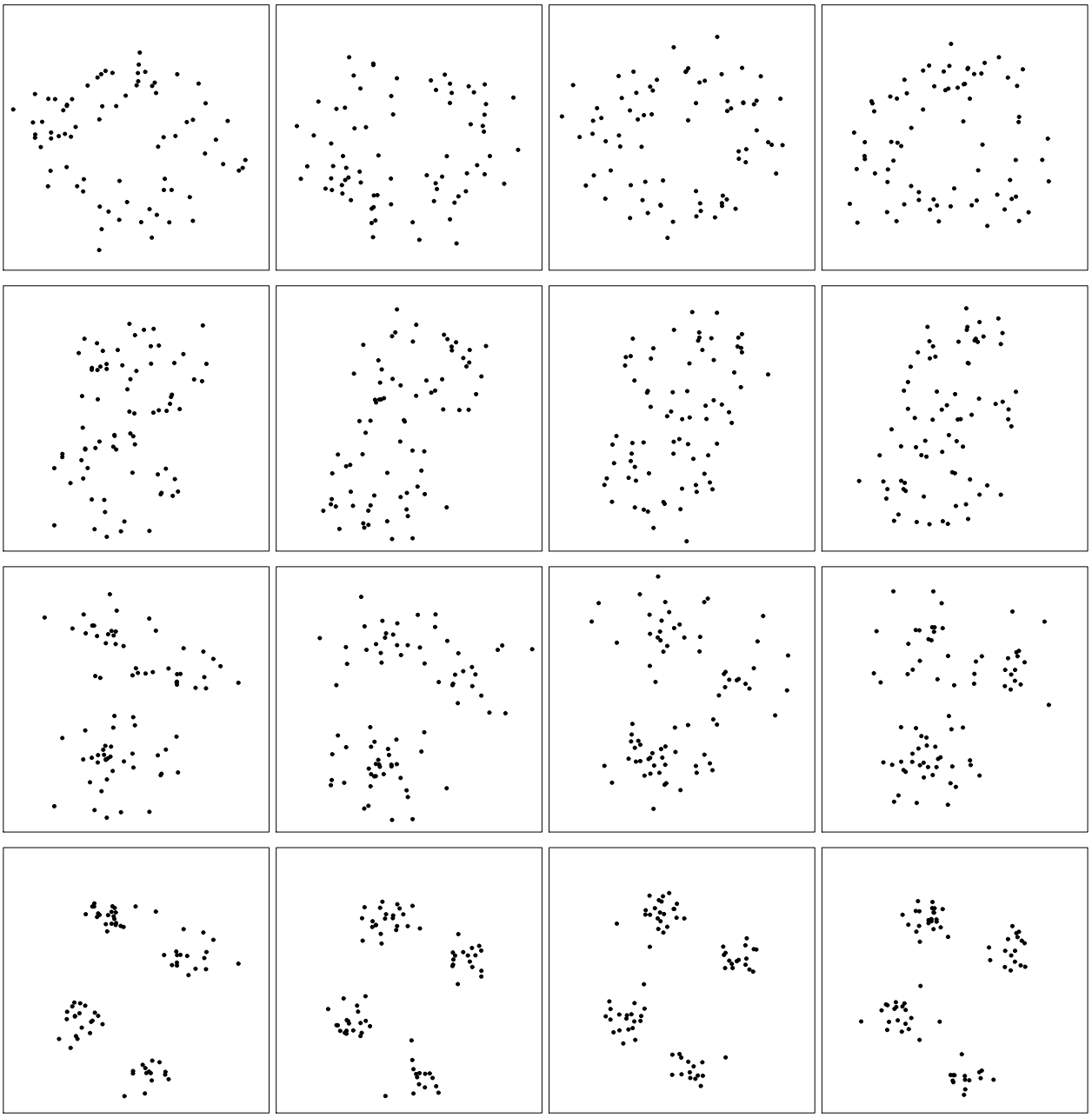
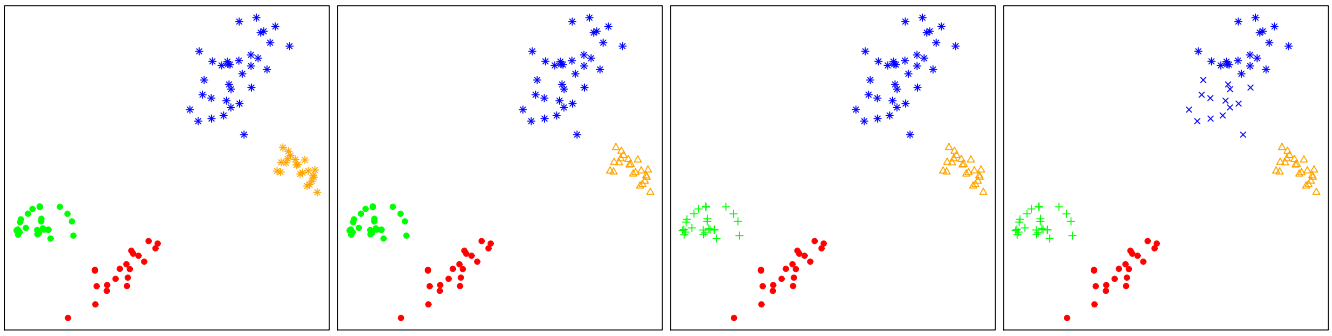


Figure S-2. Ruspini dataset: datasets resampled with the proposed method for spherical clusters under the null hypotheses $H_0 : K = 1$ *(first row),* $H_0 : K = 2$ *(second row),* $H_0 : K = 3$ *(third row), and* $H_0 : K = 4$ *(last row).*

observations in reduced space (Khattree and Naik, 2002). A different approach to representing multivariate data is provided by Andrews (1972) who suggested representing each multivariate observation in terms of a curve, much like a Fourier transform. Khattree and Naik (2002) contend that Andrews' curves are unique in that they alone have some mathematical justification in their construction, besides being also known to possess some invariance properties. In doing so, they go a considerable distance in removing the ambiguities that arise from representing multivariate observations using methodologies such as trellis plots, parallel coordinate plots, biplots, etc where ordering of the observations may end up impacting visual interpretation. Note that Andrews (1972) postulated a set of properties governing such curves and provided two possible functional relationships that could be used. Other Andrews curve representations have been provided by Kulkarni and Paranjpe (1984), Tukey – as related in Gnanadesikan (1997) –

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.



(a) (b) (c) (d)

Figure S-3. Simulated dataset: Model-based clustering solutions obtained for (a) $K = 2$, (b) $K = 3$, (c) $K = 4$, and (d) $K = 5$. Colors represent true clusterings while characters represent the predicted groupings.

and Wegman and Shen (1993). A particularly useful variant was provided by Khattree and Naik (2002) who defined their version of the Andrews' curve for a multi-dimensional observation $\boldsymbol{x} = (x_1, x_2, \ldots)'$ to be given by the function

$$g_{\boldsymbol{x}}(t) = \frac{1}{\sqrt{2}} \{x_1 + x_2 \left[sin(t) + cos(t)\right] + x_3 \left[sin(t) - cos(t)\right] + x_4 \left[sin(2t) + cos(2t)\right] + x_5 \left[sin(2t) - cos(2t)\right] + \ldots\}, \qquad -\pi \leq t \leq \pi.$$

The advantage of this particular functional representation over the others is that $g_{\boldsymbol{x}}(t)$ is more faithfully descriptive in the sense that data variation is less mixed with wave variation than that in the other Andrews' functions (Khattree and Naik, 2002). We use this particular representation for our displays in Section 3.2.

S-3.2 Clustering the Iris dataset using the k-means algorithm

Figure S-5a provides a display of the Iris' dataset with the true class (species) identities. Figures S-5b-h display partitionings of the Iris' dataset obtained using the k-means algorithm using different numbers of clusters K, along with the quality of each grouping relative to the true in terms of $\mathcal{R}$. The indifferent quality of the partitionings using k-means is reflective of the well-known fact that each of the Iris' species have very dissimilar dispersion structures, in both magnitude and orientation. The assumption of homogeneous spherical clusters, as required by the k-means algorithms often results in grossly incorrect partitionings, as seen by the Andrews curves in Figure S-5.

S-4. EXPERIMENTAL EVALUATIONS

Figures S-6, S-7, and S-8 illustrate the results summarized in Tables 1,2,3 respectively. The figures provide box-plots on the estimated $\hat{K}$ obtained using each of the methods being evaluated as well as of their $\mathcal{R}$-deficiencies $\mathcal{D}$. Overall, we note that the proposed bootstrap procedure outperforms both versions of the gap statistic as well as the method of Krzanowski and Lai (1985) and the approach based on BIC for the k-means and model-based clustering solutions, respectively.

S-4.1 Time Taken

Table S-1 provides the average times needed to construct a quantitation map for one dataset for our experiments with $n = 500$. We note that our entire methodology can be parallelized, however, we have not used parallel processing in our calculations or in reporting these computation times.

S-5. ILLUSTRATION ON A DATASET WITH NON-ELLIPSOIDAL CLUSTERS

We finally investigate the possibility of using our dataset in clustering within the framework of non-ellipsoidal clusters. We used the Clustering Algorithms Referee Package (CARP) of Melnykov and Maitra (2011) to simulate a

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

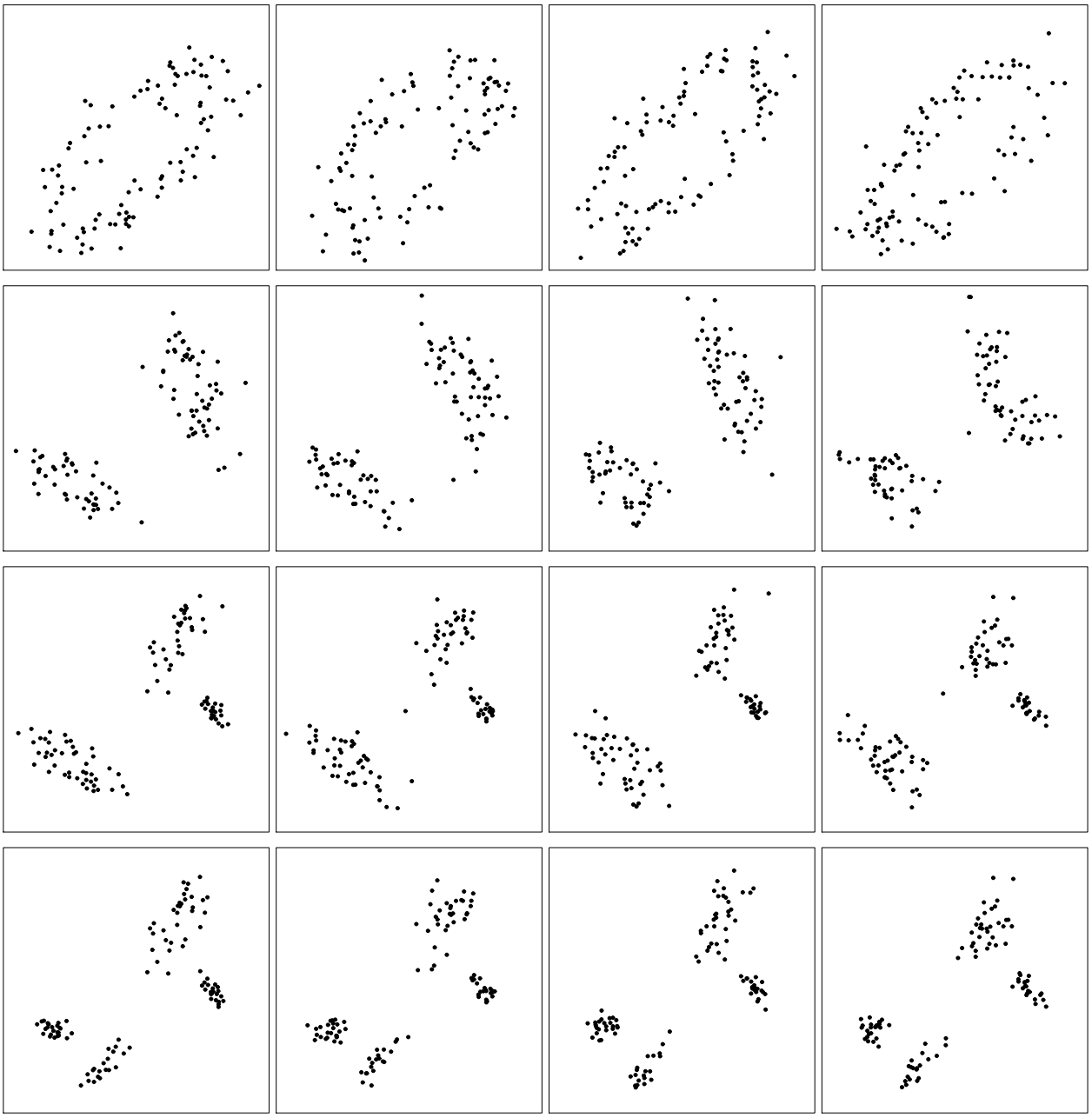
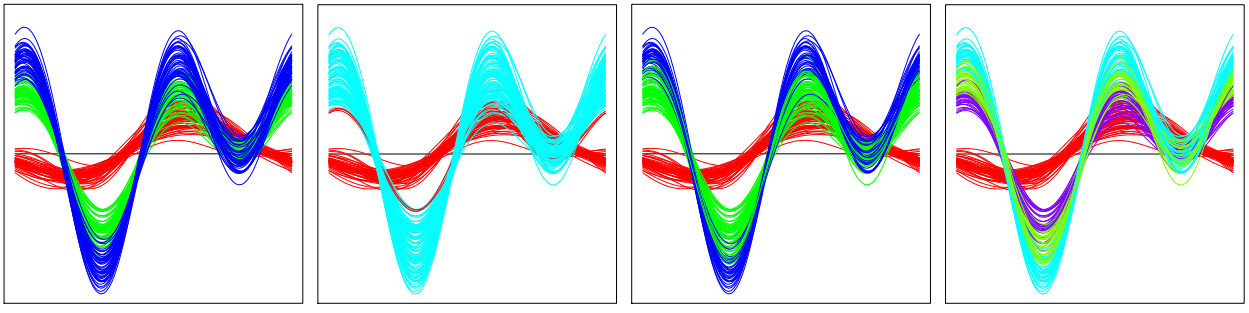


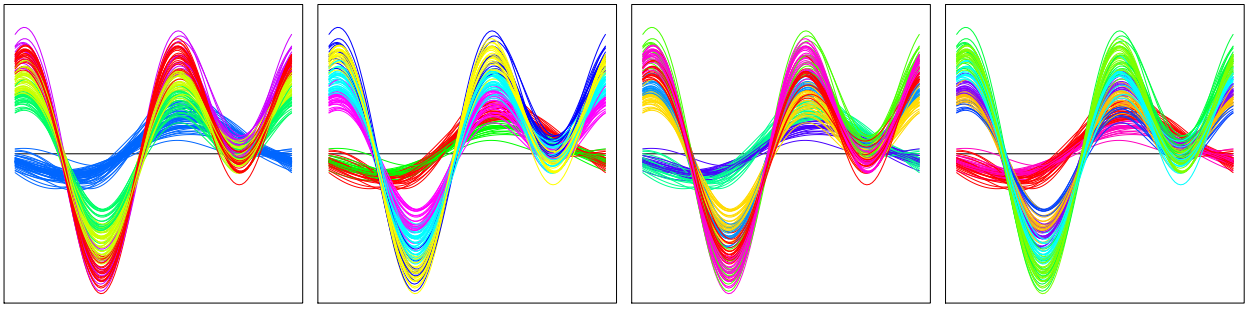
Figure S-4. Simulated dataset: datasets resampled with the proposed method for general clusters under the null hypotheses $H_0 : K = 1$ (first row), $H_0 : K = 2$ (second row), $H_0 : K = 3$ (third row), and $H_0 : K = 4$ (last row).

two-dimensional dataset of 500 observations from three nonhomogeneous non-ellipsoidal clusters using the Clustering Algorithms Referee Package (CARP) of Melnykov and Maitra (2011). This dataset is displayed in Figure S-9a. We used hierarchical clustering with single linkage to parition this dataset into one through six groups. For each case, we applied a multivariate Box-Cox transform (Hernandez and Johnson, 1980; Mardia, 1980; Gnanadesikan, 1997) to the observations in each individual cluster: the result for three clusters is displayed in Figure S-9b. We then applied our methodology of Section 2.3.2 and obtained bootstrapped realizations from the null distribution for the ellipsoidal cluster case. Each observation thus obtained was then transformed back using the inverse multivariate Box-Cox transform to yield a sample from the null distribution in the original data space. Each resampled dataset was clustered using hierarchical clustering with single linkage and the q-value quantitation map of Figure S-9c was obtained. The

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.



(a) true (b) $K = 2, \mathcal{R} = 0.540$ (c) $K = 3, \mathcal{R} = 0.730$ (d) $K = 4, \mathcal{R} = 0.650$



(e) $K = 5, \mathcal{R} = 0.608$ (f) $K = 6, \mathcal{R} = 0.448$ (g) $K = 7, \mathcal{R} = 0.475$ (h) $K = 8, \mathcal{R} = 0.436$

Figure S-5. Andrews' curves of Iris data colored according to (a) the true class identities and (b–h) groupings obtained using k-means clustering with 2 through 8 groups, along with their clustering performance relative to the true as calculated by the adjusted Rand index ($\mathcal{R}$). In (a), red curves denote the I. setosa *observations, green curves are the* I. versicolor *observations while blue curves display the* I. virginica observations.

quantitation map indicates that no groups more than three significantly better describe the data. Although small-scale in scope, this example illustrates the possibility of employing our methodology when we are able to find a suitable function Ψ as mentioned in Section 2.3.2. In this case, the function Ψ was provided by the multivariate Box-Cox transform: we note however that this may not always be the case. Nevertheless it provides us with some pointers as to possible approaches that may be employed for datasets with groups that are more general in shape and structure than ellipsoidal clusters.

References

Andrews, D. F. (1972), "Plots of High-dimensional Data," *Biometrics*, 28, 125–136.

Gnanadesikan, R. (1997), *Methods for Statistical Analysis of Multivariate Observations*, New York: Wiley, 2nd ed.

Hernandez, F. and Johnson, R. A. (1980), "The Large-sample behavior of transformations to normality," *Journal of the American Statistical Association*, 75, 855–861.

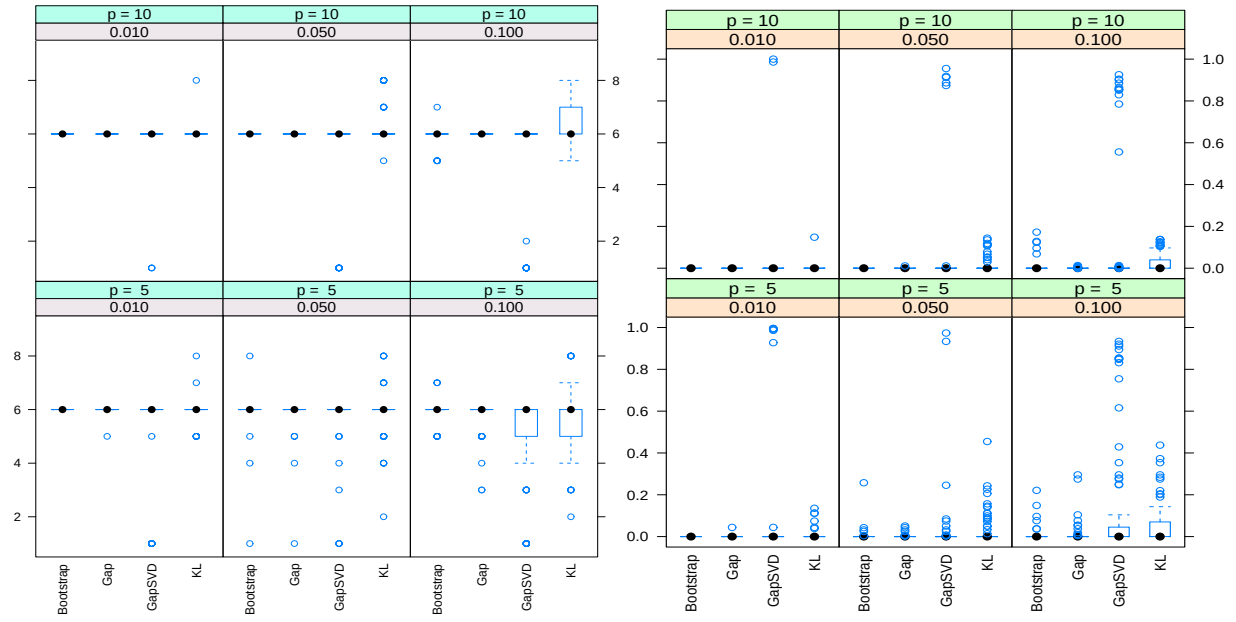
Khattree, R. and Naik, D. N. (2002), "Andrews Plots for Multivariate Data: Some New Suggestions and Applications," *Journal of Statistical Planning and Inference*, 100, 411–425.

Krzanowski, W. J. and Lai, Y. T. (1985), "A criterion for determining the number of groups in a data set using sum of squares clustering," *Biometrics*, 44, 23–34.

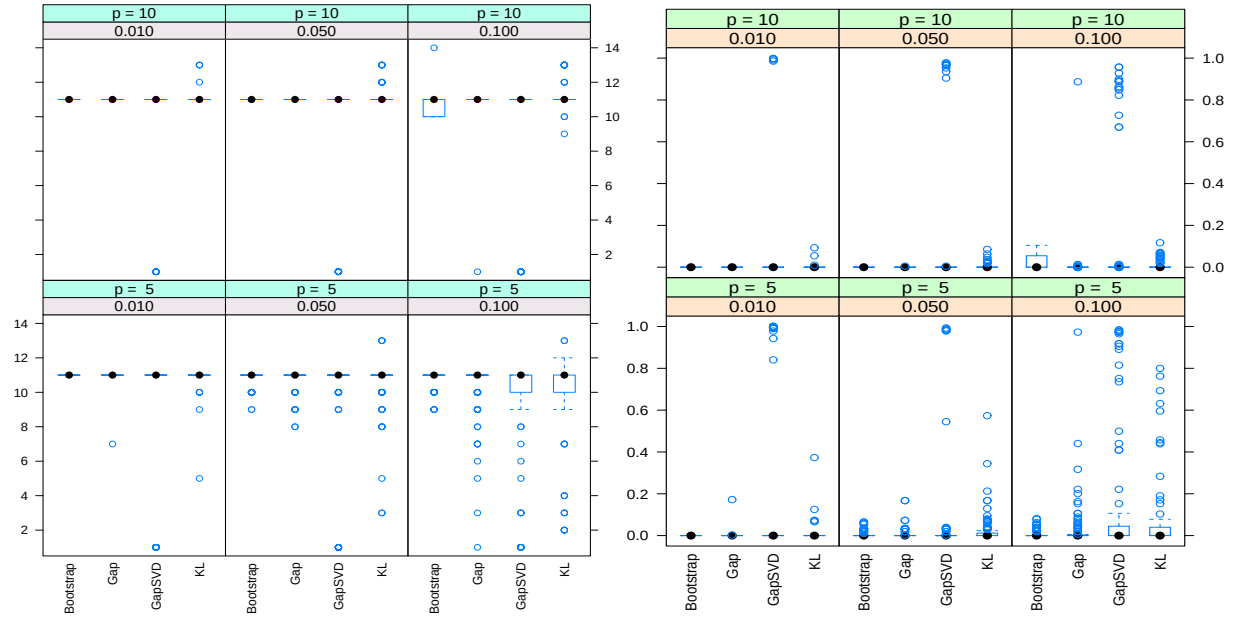
Kulkarni, S. R. and Paranjpe, S. R. (1984), "Use of Andrews' function plot technique to construct control curves for multivariate process," *Communications in Statistics: A - Theory and Methods*, 13, 2511–2533.

Mardia, K. V. (1980), "Tests of Univariate and Multivariate Normality," in *Handbook of Statistics*, ed. Krishnaiah, P. R., New York: North-Holland, vol. 1, pp. 279–320.

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.



(a) $\hat{K}$ (b) $\mathcal{D}$



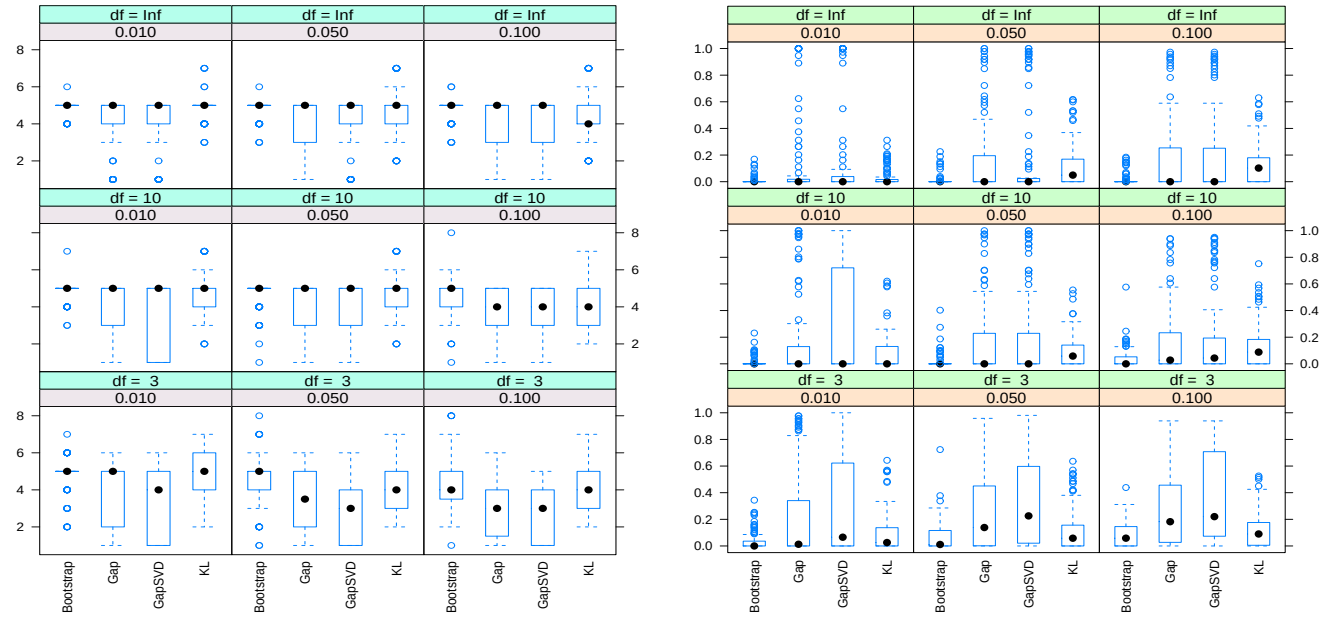
(c) $\hat{K}$ (d) $\mathcal{D}$

Figure S-6. Distribution of $\hat{K}$ and $\mathcal{D}$ for each of the methods corresponding to Table 1 for $\breve{\omega} = 0.01, 0.05, 0.1$ for true (a-b) $K = 6$ and (c-d) $K = 11$.

Melnykov, V. and Maitra, R. (2011), "CARP: Software for Fishing Out Good Clustering Algorithms," *Journal of Machine Learning Research*, 12, 69–73.

Ruspini, E. (1970), "Numerical methods for fuzzy clustering," *Information Science*, 2, 319–350.

Theus, M. (2008), "High-dimensional data visualization," in *Handbook of Data Visualization*, eds. Chen, C.-H.,



(a) $\hat{K}$ (b) $\mathcal{D}$

Figure S-7. Distribution of $\hat{K}$ and $\mathcal{D}$ for each of the methods corresponding to Table 2 for $\breve{\omega} = 0.01, 0.05, 0.1$.

Härdle, W., and Unwin, A., Santa Clara, CA, USA: Springer-Verlag TELOS, 1st ed., pp. 151–178.

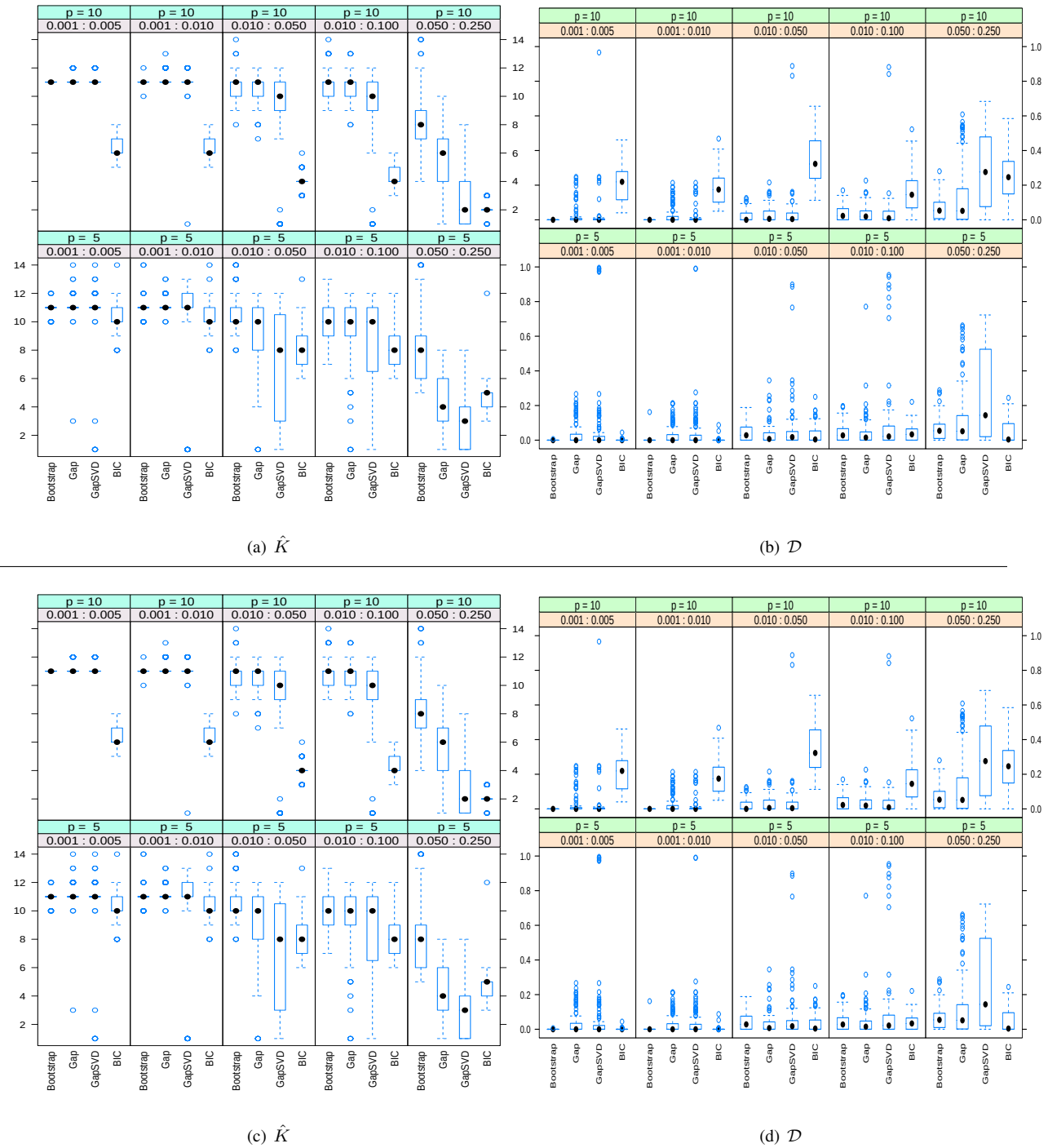
Ward, J. H. (1963), "Hierarchical grouping to optimize an objective function," *Journal of American Statistical Association*, 58, 236–244.

Wegman, E. J. and Carr, D. B. (1993), "Statistical Graphics and Visualization," in *Handbook of Statistics*, ed. Rao, C. R., Amsterdam: North-Holland, vol. 9, pp. 857–958.

Wegman, E. J., Carr, D. B., and Luo, Q. (1993), "Visualizing multivariate data," in *Multivariate Analysis, Future Directions*, ed. Rao, C. R., Amsterdam: North-Holland, vol. 9, pp. 423–466.

Wegman, E. J. and Shen, J. (1993), "Three-dimensional Andrews Plots and the grand tour," in *Computing Science and Statistics: Proceedings of the 24th Symposium on the Interface San Diego*, pp. 284–288.

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.



(a) $\hat{K}$ (b) $\mathcal{D}$

(c) $\hat{K}$ (d) $\mathcal{D}$

Figure S-8. Distribution of $\hat{K}$ and $\mathcal{D}$ for each of the methods corresponding to Table 1 for $\bar{\omega}$:$\breve{\omega}$ = 0.001:0.005, 0.001:0.01, 0.010:0.05, 0.01:0.1, 0.05:0.25, for true (a-b) $K = 6$ and (c-d) $K = 11$.

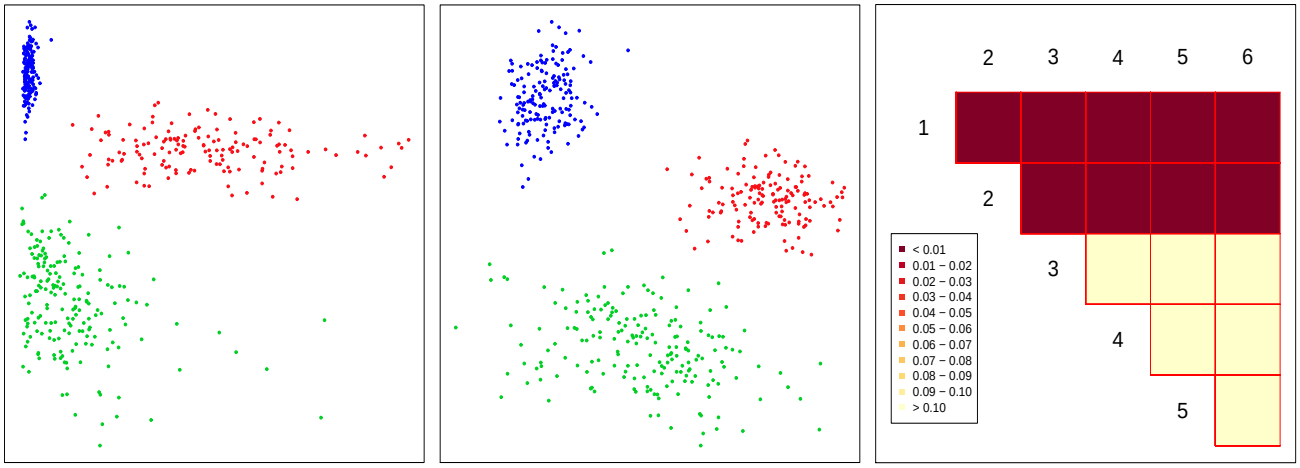
This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.

Table S-1. Average time needed to construct a quantitation map for the case of (top block) spherical clusters, (middle block) heavier-tailed spherical clusters and (lower block) general ellipsoidal clusters.

$\check{\omega}$	$K = 6, p = 5$	$K = 6, p = 10$	$K = 11, p = 5$	$K = 11, p = 10$
0.01	25.166s	30.805s	6m 17.594s	6m 42.154s
0.05	26.136s	29.819s	6m 23.264s	6m 42.269s
0.1	26.2s	30.912s	5m 57.463s	6m 24.834s

$\check{\omega}$	$\nu = 3$	$\nu = 10$	$\nu = \infty$
0.01	7.866s	7.553s	7.447s
0.05	7.912s	7.882s	7.463s
0.1	7.988s	7.780s	7.488s

$\bar{\omega} : \check{\omega}$	$K = 6, p = 5$	$K = 6, p = 10$	$K = 11, p = 5$	$K = 11, p = 10$
0.001 : 0.005	2m 11.039s	2m 30.157s	32m 40.601s	37m 30.292s
0.001 : 0.01	2m 7.475s	2m 29.791s	35m 39.282s	37m 34.564s
0.01 : 0.05	2m 9.504s	2m 32.964s	35m 45.787s	38m 25.425s
0.01 : 0.1	2m 8.422s	2m 31.096s	34m 17.423s	35m 18.247s
0.05 : 0.25	2m 11.989s	2m 35.338s	36m 0.167s	35m 37.949s



(a) Simulated dataset (b) Box-Cox transformed dataset (c) q-value map

Figure S-9. Example 2. Performance of the proposed procedure for non-Gaussian data.

This is an Accepted Manuscript of an article published by Taylor & Francis in Journal of the American Statistical Association on January 30, 2012, available online: http://www.tandf.com/10.1080/01621459.2011.646935.